Awareness, Control and Responsibility for Implicit Bias
The Continuum Thesis

Stammers, Sophie

Awarding institution:
King's College London

AWARENESS, CONTROL AND RESPONSIBILITY FOR IMPLICIT BIAS:

THE CONTINUUM THESIS

Thesis by

Sophie Stammers

Submitted for PhD

in Philosophy

KING'S COLLEGE LONDON

2016

ABSTRACT

A growing body of empirical research reveals that implicit biases manifest in many of our actions. It has been suggested that there is a fundamental distinction between (i) our implicit biases and the actions which they influence; and (ii) attitudes such as beliefs that we attribute to persons and think of as agential, and the actions that they guide. Call these the 'substantial distinction' (SD) views.

According to SD arguments, implicit biases are distinguishable from beliefs and other agential attitudes, on the basis of one or more of the following features:

- We lack awareness of our implicit biases.
- Implicit biases are associative, and so they lack the appropriate structure to enter into logical inference relations with mental states that have propositional content.
- We lack control over the formation of our implicit biases, and over the execution of our implicitly biased actions.

Some SD theorists have further argued that because implicit biases and implicitly biased actions lack one or more of the above features, they are not appropriate candidates for normative evaluation, and we are therefore not morally responsible for our implicitly biased actions.

I reject the central claim of the SD view, namely, that there is a fundamental distinction between implicit biases and agential attitudes such as beliefs, and the actions guided by each. I argue that at least some of our implicit biases are propositional in structure, and that we have the same kind of awareness and control of at least some of them, and the actions that they guide, as we do of at least some of our beliefs, and belief-guided actions. As a result, there is no principled way in which to maintain the required substantial distinction. Having shown that the SD view fails, I develop a 'continuum thesis' on which implicit biases and beliefs are not fundamentally discontinuous, and at least some of the former share all of their characteristics with at least some of the latter. I argue that this account is best able to accommodate the findings on implicit bias. According to the continuum thesis, we have a sufficient level of awareness and control such that at least some implicit biases are agential, and at least sometimes, agents are morally responsible for implicitly biased actions.

CONTENTS

INTRODUCTION

A growing body of research reveals that individuals often make choices or perform actions which suggest that they associate negative qualities with, or that they have negatively evaluated, members of a particular social group, even though the individuals in question seem to be unaware that this is the case, and do not intend their behaviour to exhibit such disfavouring treatment. For example, people perform more quickly and accurately on laboratory tasks when matching concepts in accordance with a social stereotype (such as when matching concepts denoting men with those denoting career achievement; and those denoting women and with the notions of family and caregiving) as opposed to when matching concepts counter to a social stereotype. Further, results on these laboratory matching tasks correlate with real-world discriminatory actions, such as subtly hostile behaviour (in body language and demeanour, for instance) towards members of particular racial groups in certain social interactions; as well as preferential treatment of some social groups in more deliberative scenarios, such as when evaluating C.V.s, hiring candidates or prescribing medication.[1] Whilst a number of different psychological mechanisms may be implicated in guiding the different actions mentioned above, psychologists and philosophers alike have labelled such actions as cases of 'implicit bias', or, more precisely, as actions which manifest implicit bias.

The term 'implicit bias' is often used to refer to an attitude which guides an action. However, it is also sometimes used to refer to a specific episode of biased behaviour such as the following: 'His hiring of Eleanor over Asha, who is clearly more skilled, was a case of implicit bias', or 'Dan is concerned because of the implicit bias in Rosie's marking', both of which might be understood as referring to actions which have been guided, at least in part, by a biased attitude. To avoid confusion, I will reserve the term 'implicit bias' for referring to the first sense, to the attitude, and will say that an action can *be* implicitly biased, and that

[1] For results on laboratory matching tasks, see: Greenwald, *et al.*, 1998; Nosek and Banaji, 2001; Nosek *et al.*, 2007; for subtly differential treatment in social interactions see Chen & Bargh, 1997; Dovidio, *et al.*, 1997; McConnell & Leibold, 2001; Wilson *et al.*, 2000; and for preferential treatment in more deliberative scenarios see Ulhmann & Cohen, 2005; Green, *et al.*, 2007; Rooth, 2007. I will give a more thorough overview of these results in Chapter 1.

implicit bias may 'manifest in an action', if that action is influenced by an implicitly biased attitude.

There is disagreement among scientists, and among philosophers, as to exactly what is 'implicit' about implicit bias. I will outline the background to the empirical definition in Chapter 1, §1.2. Further to this, a number of philosophers have argued that implicit biases are a *sui generis* kind of state, or, at least, that they are fundamentally different in kind to more familiar attitudes such as beliefs which we tend to attribute to persons, and to think of as agential, insofar as the latter sort of attitudes are part of the agent's evaluative perspective.[2] This sort of argument is proposed on the basis that implicit biases appear to lack particular properties that are characteristic of beliefs. These philosophers point out, for example, that we seem to be unaware of our implicit biases, or that we seem to be unable to control their influence on our actions. These claims are examples of what I will refer to as the 'substantial distinction' view (or 'SD' view for short). A subset of these SD theorists argue that, because of the ways in which implicit biases differ from agential attitudes and actions, implicit biases, and the actions that they influence, fail to meet at least some of the criteria necessary for moral responsibility, and so we cannot be morally responsible for having implicit biases, or for manifesting implicit bias in our actions. I will refer to this subset of views as the 'Substantial Distinction: Responsibility' views, or 'SDR' views for short.

One might think that something like the SD, and the related SDR views are an intuitive interpretation of a set of surprising and unsettling empirical findings. My aim in this thesis, however, is to show that we are unable to provide an account which picks out the class of implicit biases and shows them to be substantially distinct from the class of agential attitudes, such as beliefs: any account that is broad enough to include all cases of implicit bias, and the actions that they influence, will be too permissive to do the required SD work, and will also include at least some agential attitudes, such as beliefs, and the actions that they guide. I argue that there is no way to draw a principled distinction between implicit biases, and the actions that they influence, on one hand, and agential attitudes such as beliefs, and the actions that they guide, on the other. As such, any construal of the SD account will fail, and since the SDR account depends on the SD account, with the failure of the latter will come the failure of the former.

---

[2] I will characterise the term 'agential' and its cognates more fully in Chapter 2.

For the time being, then, talk of 'implicit biases' is intended to pick out the set of attitudes which are the subject of a number of empirical studies (as we shall see in Chapter 1), where participants are shown to associate particular social groups with certain stereotypical traits, without implying a deeper ontological commitment to a distinctive kind of attitude.

The argument will proceed as follows: In Chapter 1, I present a summary of evidence of implicit bias from cognitive science. There, I outline the main research paradigms and and offer a working definition of the phenomenon. In Chapter 2, I present the accounts of five philosophers who argue that there is a substantial distinction between our implicit biases and the actions which they influence, and agential attitudes such as beliefs, and the actions which they guide. The philosophical accounts that I shall focus on are those of Daniel Kelly and Erica Roedder (2008); Tamar Szabó Gendler (2008a, 2008b); Jennifer Saul (2013); and Neil Levy (2013, 2014a, 2014b). These are the main proponents of the SD view of implicit bias that I mentioned above. I will also outline the 'substantial distinction: responsibility' views or 'SDR' views in Chapter 2. The SD and SDR arguments share a number of common themes concerning

(1) awareness of the attitudes in question, and how they guide actions;
(2) the structure and processing of attitudes; and
(3) control over the attitudes and the actions that they guide.

In response to the SD and SDR views, I will defend a 'continuum thesis' of implicit bias, according to which, and contrary to the SD and SDR claims, there is in fact no single characteristic that all beliefs, and belief-guided actions have that all implicit biases, and implicitly biased actions fail to have. I address SD(R) claims made on the basis of characteristics (1)-(3) over the following three chapters.

In Chapter 3, I consider arguments that focus on (1) as above, and argue that there is no substantial distinction between (i) implicit biases, and the actions that they influence; and (ii) agential attitudes such as beliefs, and the actions that they guide, on the basis of the kind of awareness that we have of each. Following Jules Holroyd (2015), I argue that we have as much inferential and observational awareness of at least some of our implicit biases, and of their influence on our actions, as we do of at least some of our beliefs, and of their guidance of our

actions. I suggest that if we adopt Christina Borgoni's (2015) 'ordinary' notion of introspective awareness, we count as introspectively aware of at least some of our implicit biases and their influence on our actions. Holroyd (2015) maintains that it is not possible to be introspectively aware of the influence of implicit bias on action, but argues that there are cases of *agential* actions in which people are not introspectively aware of the attitudes that guide such actions. Holroyd's point would seem to count against the SD theory, but it is open to an objection articulated by Levy in his (2014a) account: Succinctly, agents who act without *occurrent* introspective awareness of the guiding role played by their agential attitudes may nevertheless still be able to 'effortlessly recall' these attitudes—that is, become occurrently introspectively aware of them in the presence of an 'ordinary cue'.[3] The same is not true in the case of implicit bias, and so, if Levy's account is correct, this would appear to reinstate the substantial distinction between implicit biases and beliefs on the basis of introspective awareness. In response to Levy, I argue that effortless recall of the attitudes which guide an action in the presence of an ordinary cue for those attitudes is *not* a necessary condition for an action to be agential. Accordingly, I show that we have the same kind of awareness of at least some of our implicit biases and their influence on our actions as we have of at least some of our beliefs and their influence on our actions. This stands, even if we don't assume Borgoni's ordinary notion of introspection.

The chapter also contains a positive account of the awareness that we have of at least some of our implicit biases. In §3.3, I introduce the notion of an 'observable class preference'. An agent has an observable class preference when they (i) have made multiple evaluations of some entities, evaluations of which they are introspectively aware, and (ii) the entities in question belong to the same class, and are evaluated to have qualities of the same kind and valence. I give four examples of everyday observable class preferences, to demonstrate that this is a common phenomenon. I then argue that at least some implicit biases are observable class preferences, and that we have the same kind of awareness of these implicit biases as we do of at least some of our everyday agential observable class preferences.

---

[3] I will say more about Levy's (2014a) notion of an 'ordinary cue' in Chapter 2, and offer a more critical exposition in Chapter 3.

This outcome of the argument against the SD view—that there is no substantial distinction between implicit biases and beliefs on the basis of awareness—enables us to refute the related SDR claims. Contra the relevant SDR views, I argue that if it turns out that we do lack moral responsibility for our implicit biases, and their influence on our actions, it will not be because we lack awareness of them, but because of some other distinguishing feature.

In Chapter 4, I consider arguments that focus on (2) above: the structure and processing of attitudes. I argue that there is no substantial distinction between (i) the structure of implicit biases, and the way in which they are processed; and (ii) the structure of beliefs (as an example of agential attitudes), and the way in which they are processed. I first outline the psychological theory that underpins the SD argument here, namely dual process theory. According to dual process theory, the mind is comprised of two systems, (i) the propositional system, which is sensitive to relational information; and (ii) the associative system which is sensitive only to the frequency with which a person has witnessed two things together. Supposedly, these systems process two distinct kinds of mental entities: propositions, and associations, in fundamentally different ways. In light of these claims, the relevant SD arguments are that implicit biases and beliefs are structured and processed in fundamentally different ways: associatively in the case of the former, and propositionally in the case of the latter.

These SD claims generate two testable hypotheses. HYPOTHESIS 1 is that that implicit biases are necessarily associative. HYPOTHESIS 2 is that beliefs change in response to changes in evidence. I show that both hypotheses are false. HYPOTHESIS 1 is falsified by findings discussed both Jan de Houwer (2014) and Eric Mandelbaum (forthcoming), which show that implicit biases (and implicit social attitudes more broadly) *do* update in accordance with propositional information. By describing a number of cases in which beliefs fail to update in accordance with new evidence I then show that HYPOTHESIS 2 is false. I also consider whether the claim in HYPOTHESIS 2 should be interpreted as a normative claim, but suggest that this would not reinstate the desired substantial distinction, since it is plausible that implicit attitudes in general are governed by the same epistemic norms as beliefs.

I also consider a more recent argument from Levy (2015) in which he acknowledges that implicit attitudes may be sensitive to propositional information, and may feature in some inferential transitions. Despite this

concession, Levy still upholds an SD view that implicit biases are distinguishable from beliefs, in virtue of his argument that beliefs are 'inferentially promiscuous': featuring in a much broader range of inferential transitions, and sensitive to a greater range of evidence, than implicit biases. I demonstrate that Levy's argument fails because there are some beliefs, in particular, some explicit prejudices, which are evidence sensitive to a lower degree than the most evidence sensitive implicit attitudes. Consequently, at least some of our implicit biases are propositional in structure, and feature in evidence-sensitive inferential transitions in the same way that many beliefs do.

It may seem that if implicit biases were associatively structured, and failed to be evidence sensitive, then this would rule out moral responsibility for having or acting on them, as SDR theorists maintain. Following the argument against the SD view in this chapter, I show that, if it turns out that we do lack moral responsibility for our implicit biases, and their influence on our actions, then it will not be because of their structure and the way in which they are processed.

Chapter 5 is concerned with claim (3) as above, the claim that implicit biases and beliefs (and the actions guided by each) may be distinguished by the kind of control that we exert over each. I first provide an overview of the main notions of agential control that appear in the philosophical literature, which will be relevant to the dialectic: (i) voluntary control; (ii) reasons responsiveness; and (iii) deep self accounts; as well as of some important distinctions in this context: (a) direct *vs.* indirect control; (b) initiation *vs.* intervention control; and (c) deliberative *vs.* non-deliberative control. I discuss SD claims on the basis of the control that we exert over the acquisition and update of each sort of *attitude* (§5.2) before moving to a discussion of control over the relevant *actions* (§5.3), and show, for each, that there is no way to maintain a strong distinction account.

As regards the discussion of attitudes, I show that if one thinks that we exert indirect voluntary control over belief acquisition and update, then, following a number of empirical findings, implicit bias acquisition and update can also sometimes be indirectly voluntary. I also show that if one is committed to direct doxastic control, then one is also committed to direct control of the acquisition and update of at least some implicit biases. As regards action control, I demonstrate that three distinct strategies are effective for controlling implicitly biased actions, which are also the *only* strategies available to us for controlling at least some of our everyday agential actions: (i) indirect, intervention control; (ii)

deliberative, direct, intervention control; and (iii) non-deliberative, direct, intervention control. Consequently, the SD theorist's claim that there is a kind of control that we have over all of our agential attitudes, and agential actions; that we do not have over our implicit biases, and implicitly biased actions, is false.

I subsequently consider SDR claims on the basis of control. There, I also consider how awareness interacts with the control that we exert over implicitly biased attitudes and actions, should the (apparent) lack of awareness *and* the (apparent) lack of control be jointly sufficient support for the SDR claim that we lack moral responsibility for implicit biases and the actions that they guide. I argue that SDR claims on the basis of both awareness and control are insufficient to uphold the desired substantial distinction.

These three chapters rule out all of the supposedly distinguishing features that were introduced in Chapter 2 as able to uphold a substantial distinction between implicit biases and agential attitudes, and the actions associated with each. In the final chapter, Chapter 6, I address the question of whether, following the arguments of Chapters 3-5, implicit biases, and the actions that they guide, implicate us as agents after all. I outline why I think it is plausible that, at least sometimes, we are the proper agential subjects of, and are sometimes morally responsible for, our implicitly biased actions.

Given the evidence surveyed in the previous chapters, I propose that the best account of the relationship between implicit biases and beliefs is one on which both sorts of attitudes are ordered along a continuum. At one extreme end of this continuum, we may find some beliefs which are effortlessly accessible to introspective awareness, and some actions under our immediate, direct control; whilst, at the other extreme end, we may find attitudes which are extremely difficult to introspect, and actions which may require considerable effort before they are amenable to any kind of control. However, in the middle of the continuum, there is a considerable area of overlap in which we find both a significant number of implicit biases (and implicitly biased actions) as well as many beliefs, (and belief-guided actions). To the extent that we hold agents to account, and praise or blame them for agential actions which lie in the overlap zone of the continuum, it is also appropriate to hold agents morally responsible for the implicitly biased actions which populate the same section of the continuum.

Any effort to try to save the SD(R) account by insisting that those beliefs which populate the same region of the continuum as implicit biases are *not* agential after all considerably restricts the account of human agency, because, as I demonstrate, we end up having to accept that a significant set of activities, some of which epitomise human flourishing, are not agential after all. And this, I show, commits us to an unsatisfactory and incomplete picture of human agency. As such, the continuum thesis remains the account that is best able to accommodate the findings on implicit bias, and further grants that at least some implicit biases are agential, and at least sometimes, agents are morally responsible for implicitly biased actions.

CHAPTER 1: THE EVIDENCE

This thesis examines the nature of implicit biases, whether they are substantially distinct in kind from the cognitive phenomena to which we are already committed (such as beliefs), and whether we are morally responsible for the actions that they influence. In order to do this, we need an understanding of the empirical evidence for the existence of implicit biases. To this end, this chapter summarises a range of results from cognitive science regarding implicit attitudes, and in particular, implicit biases and their manifestation in action.

In what follows, I provide an overview of the paradigms in which implicit attitudes are measured (§1.1) and a summary of the evidence which reveals that implicit biases are pervasive, and manifest in many everyday actions (§1.2). I then provide a brief elucidation of the apparent implicit/explicit distinction as it has been drawn by psychologists (§1.3)—although the question of whether this distinction picks out characteristics that all explicit attitudes have, which all implicit attitudes fail to have, or vice versa, will be the subject of later chapters. As I reject that there is a substantial distinction between implicit biases and attitudes which have been identified by my opponents as *explicit* (such as beliefs) I will also explain how I intend to use the term 'implicit' in the argument to come.

1.1. HISTORY AND DEVELOPMENT OF IMPLICIT ATTITUDE MEASURES

Experimental measurement of implicit social bias follows findings in two research streams that focus (1) on how concepts are classified in, and retrieved from, memory, and (2) on the conditions under which mental processing occurs without attention (Payne & Gawronski, 2010). A significant finding within the first stream is the 1970s discovery (Meyer & Schvaneveldt, 1971; Collins & Loftus, 1975; Neely 1977) that presenting a person with a word facilitates their access to other words which are conceptually related to the first, enabling them to more quickly access and employ such words on various lexical classification tasks. For instance, Meyer and Schvaneveldt (1971) demonstrated that when English speaking participants are tasked with indicating whether a string of letters forms an English word under time pressure, they are quicker to recognise, for example, 'butter' as an English word when first presented with the word 'bread', compared to a word that is unrelated to 'butter', such as 'window' or 'doctor'. Results

gained from these lexical decision-making tasks and other experiments in similar paradigms led to the proposal that activation of a mental entity tends to co-activate other entities which are closely associated in long-term memory (Meyer & Schvaneveldt, 1976; Anderson, 1983).

Meanwhile, in research stream (2), psychologists investigated the ways that stimuli to which participants were not directing their attention could nevertheless affect mental processing (Shiffrin & Schneider, 1977; Posner & Snyder, 1975). For instance, Shiffrin and Schneider (1977) observed that when participants were extensively trained to pay attention to a target stimulus, and that target stimulus was later displayed in the corner of the screen during another task, participants performed more poorly on the task. This was not the case for symbols participants had not been trained to pay attention to. This, and related results, led to widespread acceptance of the view that, whilst some mental processing is initiated by an intention, and requires effort on the part of the subject, other processing can occur despite the absence of intention and effortful control. Processing with these latter properties acquired the adjective 'implicit', with the former being referred to as 'explicit'. I will say more about the implicit *vs.* explicit distinction in §1.3, but it is perhaps worth pointing out now, after Payne and Gawronski (2010), that in the early days of research into implicit and explicit processes, reference to conscious awareness was not a vital part of the distinction.

These findings set the stage for broader research programmes into so-called 'implicit social cognition', exploring how people memorise, access and process cultural stereotypes: the coupling of social group concepts and traits, such as that of men with career concepts, as I mentioned in the introduction. Psychologists have referred to such stereotypes as items of cultural "knowledge" (Devine, *et al*., 2002: 846). However, given that stereotypes are not necessarily truth-tracking, this definition presents problems for philosophers, who almost universally think that knowledge is factive. This is not to say that psychologists are committed to the idea that stereotypes are founded on truth—just that the word 'knowledge' means something slightly different in each discipline. I take a 'stereotype' to be a publicly available conceptual construct in which some people, who are perceived to belong to a particular social group, are assumed to have one or more character traits—traits which may have a positive or negative valence. I

stress that, on this understanding, stereotypes do not have to track truth, or to even be founded in some modicum of truth.[4]

In 1983, Gaertner and McLaughlin used a lexical decision-making task inspired by Meyer and Schvaneveldt's (1971) paradigm to investigate whether words denoting racial categories facilitated response times to stereotype congruent concepts. Participants in Gaertner and McLaughlin's experiment had to determine whether a pair of words were English words or nonsense words. It turned out that white participants were significantly faster to correctly identify English word pairs featuring terms denoting positive social traits (such as 'ambitious' or 'smart') when the other word was 'White' as compared with pairs featuring a positive social trait and the word 'Black' or 'Negro'. In line with their aforementioned model, (as well as that of Anderson, 1983) Gaertner and McLaughlin suggested that this indicated that positive social traits are more closely associated in long term memory with the category 'White' than with the category 'Black or Negro'.[5]

Working in a slightly different paradigm, Dovidio, Evans and Tyler (1986) found a significant relationship between the word 'black' and the speed at which white participants access negative social traits. Specifically, they discovered that when participants were presented with the words 'black' or 'white', and then asked to indicate whether a given trait 'could ever be true' of these words, white participants were reliably faster to identify negative social traits (such as 'lazy') as true of the category 'black' than of the category 'white'.[6] Participants were also faster to indicate that positive social traits could be true of the category 'white'

---

[4] For example, to observe that, in some cultures, characteristics such as 'hysterical' and 'weak' are assumed to be had by (people perceived to be) women, is not to say that therefore, in those cultures, it is generally true of the people in question that they *are* both hysterical and weak. The reasons for which particular character traits become associated with particular social groups is a valuable research avenue that likely involves probing the (often appalling) historical and structural injustices in social and economic power but I do not have space to explore this here. For brevity, I will talk about 'stereotype congruence' and 'stereotype incongruence', but I am not thereby committed to there being anything essential or revealing about the character traits predicated of people who (are perceived to) belong to particular social groups according to the dominant social stereotypes of any one culture, and do not look upon these stereotypes uncritically.

[5] Whilst 'Negro' is clearly a racialised term, one may reasonably question whether the terms 'White' and 'Black' definitely activate *racial* categories as opposed to metaphorical concepts denoting positivity and negativity in Gaertner and McLaughlin's (1983) results.

[6] The concern that 'black' and 'white' activate metaphorical concepts denoting negativity and positivity, rather than racial concepts, is also valid here. In isolation of other corroborating results, whether the associations observed in these experiments hold between racial or metaphorical concepts remains open. However, the introduction of laboratory tasks which included *pictorial* representations of people perceived to be from different racial categories would later resolve this ambiguity, as I discuss in the main text shortly.

than of the category 'black', as was the case with participants in Gaertner and McLaughlin (1983).

The emergence of the Implicit Association Test, or 'IAT' for short, allowed researchers to measure the differential association between two target concepts and social attributes (see for example, Dovidio *et al*., 1997; Greenwald, McGhee & Schwartz, 1998; Dasgupta & Greenwald 2001; Nosek *et al.*, 2007). The IAT enabled researchers to run trials in which participants matched *pictures* of people (or names) representing two different races or genders, with evaluative tokens (positive and negative words). A typical IAT trial might require participants to focus on a screen with evaluative tokens of opposing valence (the words 'good' and 'bad' for instance) displayed in the top left and top right corners. Participants will then be presented with a series of target items—often pictures of people of different races or genders, or words evoking racialised or gendered characteristics—and instructed to 'classify' them with the evaluative tokens on either side of the screen, by pressing either a left or a right key. For instance, on one trial, participants might have to classify pictures of black people with 'good' and pictures of white people with 'bad', and then on a later trial, classify black people with bad and white people with good. The order of these critical trials is usually reversed for some subset of participants to ensure that results are not just a function of the order in which the classification tasks are completed. Participants also undergo training rounds prior to the critical trials to ensure they are able to classify examples of, for instance, racialised names with the appropriate racial categories, as well as classifying examples of evaluative tokens with the appropriate valence. This ensures that results are not just a function of a participants' familiarity with one category over another.

Participants must respond as quickly as possible, whilst doing their best to avoid errors. In light of the hypothesis that the activation of mental constructs tends to facilitate access to other constructs with which the first are closely associated (Meyer and Schvaneveldt, 1971), it was hypothesised that participants would be quicker to classify target concepts with stereotype congruent evaluative tokens than with stereotype incongruent evaluative tokens. It was discovered that participants are faster to pair racialised-as-white names with positive terms than they are to pair racialised-as-black names with positive terms, (Greenwald, McGhee and Schwartz, 1998), and faster to pair black faces with negative terms than with positive terms, (Dovidio et al., 1997). Participants are also significantly

faster to respond on stereotype congruent trials when matching gender and career concepts, gender and science concepts, and age and valenced concepts (Nosek, Greenwald and Banaji, 2005) as well as sexuality and valenced concepts (Dasgupta and Rivera, 2006) and disability and valenced concepts, (Lane, Banaji, Nosek and Greenwald, 2007).

One criticism levelled against the IAT paradigm is that because it only enables the measurement of *differential* response times, it cannot tell us about the specific content of an implicit attitude. For instance, if a person is faster to respond on IAT trials matching 'white' with 'good' than 'black with 'good', we still do not know whether this is because they associate 'white' with 'good', or because they fail to associate 'black' with 'good'. Acknowledgment of this limitation gave rise to several adjusted paradigms that are able to determine which concepts the activated association holds between. The first of these is Nosek and Banaji's (2001) Go/No-Go Association Task (GNAT), which, unlike the IAT, requires participants to respond to just one target object and evaluative category, rather than comparing response latencies between two target objects. For instance, participants must indicate whether stimuli represent either the target category in question, or a particular attribute, but refrain from responding to distractor items. If subjects are both faster and more accurate to respond when target category A is paired with good words rather than bad words, researchers infer that the underlying attitude mediating these responses is an affirmation of category A and words of positive valence. It turns out that subjects are both faster and more accurate to identify 'white' with 'good' than 'white' with 'bad', and faster to identify 'black' with 'bad' than 'black' with 'good' (Nosek & Banaji, 2001). So, facilitated stereotype-congruent responses, at least in this task, were likely mediated by an affirmation of both the positive white stereotype and of the negative black stereotype.[7]

[7] As Payne and Gawronski (2010: 8) point out, some remain suspicious that when participants respond to multiple presentations of stimuli in the block format of both the IAT and the GNAT, they may develop their own heuristic to group salient categories together, and if this was the case, then neither the IAT nor the GNAT would record stable underlying attitudes of the participants, but instead attitudes which are simply a function of the sorting task. Further paradigms correct for this worry, for instance by presenting both stereotype congruent and incongruent tasks randomly within a single block, as in De Houwer's Extrinsic Affective Simon Task (De Houwer, 2003). See also the Single-Block IAT (Teige-Mocigemba, Klauer, and Rothermund, 2008) and the Recoding-Free IAT (Rothermund, Teige-Mocigemba, Gast, and Wentura, 2009).

One might reasonably wonder whether being slower to match stereotype-incongruent concepts than stereotype-congruent concepts in an abstract experimental setting has any particular implications for behaviour outside the laboratory. It turns out that it does, as I demonstrate in the next section.

1.2. Implicit Attitude Measures Predict Biased Behaviour

Plenty of studies show that people who exhibit stereotype-congruent responses on tests like the IAT (call them people who are 'high' in IAT bias) are more likely to perform a range of stereotypical behaviours in both experimental simulations of real-world contexts, and in actual everyday decisions. I give an overview of just some of the results in the following.

Scores on implicit measures of attitude predict a range of 'microbehaviours'—subtly discriminatory behaviour. For instance, McConnell and Leibold (2001) had participants undertake a race IAT and then put them in a number of scenarios which required social interaction with black and white experimenters. It turned out that people with high stereotype-congruent responses on the IAT are more likely to exhibit signs of social unease when conversing with a black experimenter. In particular higher IAT biases "predicted greater speaking time, more smiling, more extemporaneous social comments, fewer speech errors, and fewer speech hesitations in interactions with the White (vs Black) experimenter," McConnell and Leibold, (2001).[8] Scores on implicit attitude measures also predict less eye contact and increased blinking when in conversation with a black experimenter (Dovidio *et al*., 1997), as well as predicting the number of times participants touch the hand of a black experimenter, (Wilson *et al*., 2000).

IAT bias also predicts a range of 'real-world' discriminatory behaviours. Doctors with high IAT race bias are less likely to offer treatment to black patients with the clinical presentation of heart disease than to white patients with the same clinical presentation of the disease (Green, et al., 2007). Swedish recruiters with high IAT race bias are significantly less likely to offer a job interview to an applicant with a name that they perceive as belonging to a Muslim, as compared to applicants with a Swedish name (Rooth, 2007). Judges with high IAT race bias

[8] Social interactions were observed by the experimenters themselves, as well as two trained judges via videotape. There was good agreement across all observers, (McConnell and Leibold, 2001).

gave harsher (mock) sentences to defendants when primed with information associated with black people than those who exhibited less IAT bias (Rachlinski *et al.*, 2009). White student participants with high IAT race bias are more likely to recommend disproportionate budget cuts for Jewish, Asian and Black student organisations, (Rudman and Ashmore, 2007).[9] These results were observed, even when participants indicated on questionnaires (where participants report the attitudes they take themselves to have, without the significant time pressure of the IAT and similar measures) that they have broadly egalitarian attitudes. Indeed, there is typically no or only a low correlation between IAT (and similar) measures of attitude and prejudice exhibited on self-report questionnaires, a fact which is often taken to be indicative that implicit bias is a distinct phenomenon from the kind of attitudes that people profess to have on self-reports (Nosek et al. 2007).

Even more worryingly, those required to make fast—often life-ending—decisions about whether a person is carrying a weapon are also likely to be affected by implicit bias. When participants in an experiment are pressed to respond quickly, those who see a black face before they see either a gun or a non-gun object are more likely to mistake a non-gun object for a gun than when they see a white face (Payne, 2005, 2006). This phenomenon, which Payne terms 'weapon bias', correlates with IAT bias (Payne, 2005). Weapon bias has been shown to affect US police officers in an experimental setting (Plant & Peruche, 2005).

There are a number of other experiments which show bias in behaviour, without correlating such bias with scores on an implicit attitude measure, such as the IAT, but which take other measures to demonstrate that participants do not seem to intend to make discriminatory choices. Insofar as bias seems to manifest in action automatically, these results are also generally considered to reveal cases of implicitly biased behaviour. A notable example is found in Uhlmann and Cohen (2005), who had participants assess the resumes of two candidates, and then judge who is most suited to a professional role. Uhlmann and Cohen divided participants into two groups and asked them to rate the suitability of two (made up) candidates for the role of police chief, then asked them to decide who they

[9] These participants are also more likely to report having engaged in racial discrimination in the past, including using verbal slurs, exclusion behaviours and even physical harm, although these would seem to be discriminatory behaviour of which the participants in question may well be aware.

would hire and to give their reasons as to why. In group A, candidates are presented with an educated but not very streetwise woman, and a streetwise, but not very educated man. In group B, candidates are presented with an educated but not very streetwise man, and a streetwise, but not very educated woman. On average, *both* groups preferred the man, with participants in group A citing his streetwise credentials as the desired criteria for the role, whilst group B cited their preferred candidate's educated background. The candidate descriptions were copied verbatim, differing only for the gender of the name, and so the implicitly preferred 'qualification' would seem to be neither education nor streetwise credentials—but gender. During questioning, however, participants make no reference to gender as a factor in their deliberation.

A range of other studies are often quoted in the implicit bias literature, which might be described as investigating the extent to which phenomena Uhlmann and Cohen (2005) observed in the laboratory affect actual hiring and candidate appraisal. For instance, Bertrand and Mullainathan (2004) sent a sample of C.V.s with similar qualifications to a number of US employers, which appeared to come from either 'Emily' or 'Greg' (typical racialised-as-white American names), or 'Lakisha' or 'Jamal' (typical racialised-as-black American names). It turned out that US employers are far more likely to offer job interviews to 'Emily' or 'Greg' than to 'Lakisha' or 'Jamal', despite the fact that the C.V. differed only in the names at the top. In a similar vein, Milkman (2014) showed that when university faculty are contacted by fictional prospective students seeking to discuss research opportunities, they ignore requests perceived to come from women and non-Caucasian candidates at a significantly higher rate than those perceived to come from Caucasian men, even though the content of the emails remains unchanged. Further, Budden *et al.* (2008) report that the representation of women authors in the journal Behavioural Ecology increased by 33% when they started practising anonymous submission so that editorial decisions could not be influenced by the apparent gender of the author.

Whilst these results are at least sometimes referred to as examples of implicit bias, I think that we ought to proceed with caution before concluding that the discrimination observed above by Bertrand and Mullainathan (2004), Milkman (2014) and Budden *et al.* (2008) is determinately the result of *implicit* bias. My concern with these three studies, as compared to those described previously, is that experimenters did not measure participants' self-reported

attitudes, and so these results are consistent with the possibility that *explicit* racism and sexism accounts, at least in part, for the discriminatory behaviour observed. But even leaving these three studies aside, I take it that the other results presented in this section are sufficient for it being highly likely that bias manifests in the actions of people who profess to having egalitarian attitudes—attitudes which we might think ought to prevent their discriminatory actions, and yet, which seem to fail to.

Whether there is a robust distinction in kind between the attitudes which influence action in this way and the more everyday species of cognitive phenomena to which we're already committed (such as beliefs) is one of the guiding questions of this thesis (wherein, in Chapters 3-5, I argue that no such principled distinction is defensible). In the remainder of this chapter, however, it will be useful to look at some of the ways in which the notion of the implicit has been characterised by psychologists, which will help to situate the argument to come.

1.3. How is the Notion of the Implicit Characterised?

One might wonder what it is that is supposed to be distinctly *implicit* about attitudes observed on the IAT and related paradigms, as well as the discriminatory behaviour with which these attitudes correlate. As I mentioned briefly in §1.1 (following Payne & Gawronski, 2010), the implicit/explicit distinction within cognitive science was not, at least at first, defined with respect to the subject's conscious awareness. Rather, the distinction as psychologists identified it had to do with the notion that the relevant processes are *automatic* as opposed to controlled: That is, they simply occur, without proceeding from an intention, and unfold without the guidance of attention when they influence behaviour—Schneider and Shiffrin define an automatic process as that which is activated "without the necessity for active control or attention by the subject" (1977: 2). Note that a process's being automatic in this sense is consistent with a person being aware that they are performing intention-incongruent behaviour. For instance, a patella-reflex movement is automatic, but an individual may be aware of it when it occurs. This is significant because, as I will demonstrate in Chapter 2, many philosophers (Kelly & Roedder, 2008; Saul, 2013; Levy, 2014a) suggest that unavailability to awareness is an essential characteristic of an attitude's being 'implicit'. As I will argue in more detail in Chapter 3, the fact that a process is

automatic is not incompatible with the subject's having awareness that such a process is nonetheless operative.

According to Moors *et al.* (2010: 20), it is characteristic of an automatic process that it can result in effects when the subject in question did not have the goal of achieving such an effect. So one might think that further support for the notion that the IAT measures automatic processes comes from the finding that people cannot 'fake' an IAT test: Participants do not eliminate biased responses on stereotype incongruent trials just because they are instructed to do so (Banse, Seise, & Zerbes, 2001; Steffans 2004).[10] Relatedly, those who are asked to form a goal to not stereotype are not able to reduce the extent to which their responses are stereotyped on an IAT (Lowery *et al.*, 2001).

Another way in which psychologists have distinguished implicit attitudes from explicit attitudes is with reference, not to the attitudes themselves, but with respect to the processes by which we *measure* them. (Payne & Gawronski, 2010). The IAT and related paradigms presented in §1.1 are often called 'indirect' measures of attitude. This is because on these indirect measures, attitudes are determined by requiring the participant to elicit some behaviour, other than directly telling the experimenter what their attitude is, from which an attitude may then be inferred. For example, if a person responds more quickly when pairing male concepts and career concepts than when pairing female concepts and career concepts, then it is inferred that an attitude in which male concepts are associated more closely with career concepts than are female concepts is guiding the behaviour in question. Whilst attitudes are inferred from behaviour on indirect measures, these measures demand fast responses, and participants have very little time to deliberate about how to respond, so it is thought that these measures reveal unmediated attitudes about the target items.

In comparison, asking participants to indicate whether they agree or disagree with a particular statement about a target item in a questionnaire, or to simply report their own attitudes to experimenters, are known as 'direct' measures of attitude. Here, there is no need to infer the attitude from behaviour, as it is explicitly expressed in what the participant says or writes. Whilst, as we will see in the coming chapters, many philosophers have argued that we should think that

---

[10] This result does not yet mean that participants are unable to control the manner in which implicit biases affect their behaviour in general, as I will argue more fully in Chapter 5.

the answers that people give on self-report questionnaires reveal their beliefs, it is worth being clear that, at present, there is no experimental measure which can deliver a wholly accurate, direct assessment of the statements and sentiments that participants genuinely believe: When participants respond to self-report questionnaires, they do have time to deliberate about how to respond, and so it is possible that the answers that they give are not a direct record of their attitudes, but a record of what they would *like* to be seen to believe, for instance. Research demonstrates that self report measures are open to modification when subjects are motivated to comply with social norms of prejudice (Plant & Devine, 1998; Nier, 2005). For example, participants express more highly prejudiced racial attitudes on a self-report measures when their responses are anonymous as compared to when they are reported to the experimenter (Plant & Devine, 1998). Further, if participants are led to believe that experimenters will *know* if their self-reports do not match their 'genuine' personal attitudes, (thus reducing the motivation to report socially desirable attitudes in place of accurate personal attitudes) then participants report attitudes which *do* correlate with their IAT scores (Nier, 2005).

None of this is to doubt that the experiments presented in this chapter pick out genuine discriminatory effects that are prevalent in our social interactions in the world beyond the laboratory—far from it. I think that the evidence suggests that these effects are real and pervasive. Rather, my contention is with the way in which the attitudes implicated should be characterised. So, I emphasise that when I use the term 'implicit bias', or more generally 'implicit attitude', I do not intend to refer to a *sui generis* kind of mental entity, distinguishable from beliefs on account of failing to lack a set of properties that beliefs have. Instead, I use the term 'implicit bias' to pick out the set of attitudes which are the subject of the foregoing studies, and which have been shown to manifest in all kinds of social interactions, without yet being committed to there being any particular characteristics which *all* implicit biases share, and that all beliefs, for instance, lack.

Summary

An extensive body of research reveals that people harbour what have been termed 'implicit biases', in which social groups are associated with stereotypical traits, and which appear to guide behaviour automatically, often without the subject's awareness of the guiding role that these biases play. These attitudes may be

observed on a variety of ‘indirect’ measures, typically do not correlate with the beliefs and values that people profess to hold on ‘direct’ measures, and are highly likely to manifest in many real-world decisions and actions. Whether these implicit biases are substantially distinct from more familiar cognitions such as beliefs, or whether they lie on a continuum with beliefs, is one of the guiding questions of this thesis. In the next chapter, I present a range of arguments for the former ‘substantial distinction’ view, before arguing (in Chapters 3—5) that each version of the substantial distinction view fails.

CHAPTER 2: THE SUBSTANTIAL DISTINCTION VIEW OF IMPLICIT BIAS

The chapter presents a view about the nature of implicit attitudes defended by several philosophers: Daniel Kelly and Erica Roedder (2008), Tamar Szabó Gendler (2008a, 2008b), Jennifer Saul (2013), and Neil Levy (2013, 2014a, 2014b). According to these philosophers, implicitly biased attitudes have certain characteristics concerning awareness, structure and processing, and control. These philosophers further claim that, because implicit biases have these characteristics, there is a distinction in kind between (i) our implicit biases, and the actions that they guide; and (ii) attitudes that we attribute to persons and think of as agential, such as beliefs and desires, and the actions that they guide. I call this the 'substantial distinction' view of implicit bias or 'SD' for short.

A subset of SD theorists, for instance, Kelly and Roedder, Saul, and Levy, argue that, because of the ways in which implicit biases differ from agential attitudes and actions, implicit biases and the actions that they influence, fail to meet at least some of the criteria necessary for moral responsibility, and so we cannot be morally responsible for having implicit biases, or for manifesting implicit bias in our actions. I will refer to this subset of views as the 'Substantial Distinction: Responsibility' views, or 'SDR' views for short.

I first identify some key concepts that are central to the dialectic, and explain how I shall be using them (§2.1). Following this, I present the views of the five philosophers as above who have recently written on the topic of implicit bias and who defend versions of the SD view: Jennifer Saul; Daniel Kelly and Erica Roedder; Tamar Szabó Gendler; and Neil Levy (§2.2). I then identify some common themes in the SD arguments (§2.3). Accordingly, arguments for a substantial distinction between implicit biases and agential attitudes are made on the basis of

(i) our awareness of the attitudes and how they guide actions;
(ii) the structure and processing of the attitudes; and
(iii) our control over the attitudes and the actions that they guide.

These three factors will be the focus of the three subsequent chapters.

In the summary of this chapter, I outline the account that I will present in this thesis, for which I will gradually build support in the dialectic to come. As I respond to SD arguments made on the basis of (i)-(iii) over the course of Chapters 3-5, I gradually build support for what I call the 'continuum thesis' of implicit bias. According to the continuum thesis, there is no single characteristic that all beliefs and belief-guided actions have; that all implicit biases, and implicitly biased actions lack. Because SDR views rely on the truth of SD views, the establishment of the continuum thesis gives us a response to SDR views 'for free'.

2.1 Terminological Clarifications

Before presenting the SD and SDR arguments, it will be helpful to have some background on two concepts to which the forthcoming arguments will frequently refer. These are the concepts of agency and responsibility. In what follows, I will briefly outline how these terms are to be understood in my discussion.

2.1.1. *Agency*

Much philosophy is premised on the idea that humans (and perhaps also some non-humans) are agents, beings who have the capacity to bring about changes in their bodies and beyond in accordance with attitudes such as their desires, beliefs, and intentions, although the precise details of how they do this differ across accounts.[11] There are at least some bodily changes which are considered to be uncontroversially non-agential, at least in part because they are not guided by these sorts of attitudes. For instance, both the activity of digestion and reflex movements bring about changes in the agent's body but they do so without the guidance of desires, beliefs or intentions, and so are typically considered to be non-agential occurrences.

It is widely thought that attitudes such as desires, beliefs, and intentions *involve* the agent in some important sense. Exactly what this amounts to differs across accounts. For instance, according to Neil Levy (2014a), for an attitude that *P* to be agential, the agent has to have a particular sort of awareness of *P*. According to Pamela Hieronymi (2008, 2009), for an attitude that *P* to be agential, the agent has to have 'settled the question' of whether *P*. When agential

[11] For some example accounts of agency see Davidson, 1971; Watson, 1975; Dennett, 1987; Velleman, 1992; Mele, 2003; Hornsby, 2004; Hieronymi, 2009.

attitudes bring about changes in the agent's movements, or further changes in the agent's attitudes, these changes are, at least sometimes, considered to be actions which are rightly identified with the agent in question. For instance, raising a glass of water to one's mouth is an agential action, if it is guided by attitudes which involve the agent, such as a desire to quench one's thirst, the belief that water quenches thirst, and the intention to utilise the thirst-quenching properties of the water to quench one's thirst.

My aim in this thesis is not to offer an account of agency. Instead, my aim is to investigate whether the conditions which, it has been argued, are necessary for an attitude such as a belief to be agential, as well as the conditions necessary for belief-guided actions to be agential, are also present in the case of at least some implicit biases, and implicitly biased actions. I shall approach that question by identifying three conditions that have been said to be necessary for agency: awareness, structure and processing of attitudes, and control. In §2.2 I outline and examine a number of arguments for the claim that implicit biases, and implicitly biased actions, do not fulfil one or more of the conditions which render attitudes such as beliefs, and the actions that they guide, as agential.

2.1.2. '*Responsibility'*

There is a large and diverse literature on what being responsible amounts to (for a detailed overview, see Fischer, 1999). Generally, there is agreement across most accounts that to be responsible for some episode of ϕ-ing is to be liable for praise or blame from one's community. A common theme between a number of accounts is that it is a necessary condition on an agent's being responsible for some ϕ-ing, that the ϕ-ing in question is, in some sense, agential. For instance, I cannot be held responsible for digesting my sandwich, because digestion is not an agential action. Accordingly, some (such as Wolf, 1990; Fischer & Ravizza, 1998) hold that (moral) responsibility is a matter of the agential ability to respond to (moral) reasons. If ϕ-ing conforms with or violates some moral norm, then agents are morally responsible for ϕ-ing if they are able to act in light of reasons they see there to be for ϕ-ing. Others (such as Watson, 1996; Smith, 2008; Hieronymi, 2008) suggest that (moral) responsibility is a matter of the manifestation one's agential attitudes. Here, if ϕ-ing conforms with or violates some moral norm, then agents are morally responsible for ϕ-ing if ϕ-ing flows from the agent's evaluative judgements and other agential commitments. Another suggestion is that agents are

morally responsible for ϕ-ing, if, in ϕ-ing, they are the proper subjects of particular sentiments (the 'reactive attitudes') that arise as a result of their "participation with others in interpersonal human relationships" (Strawson 1962). On this account, judgments of moral responsibility track the natural reactions that we have to the actions of other people, which are expressive of how much we care about their actions on the basis of our joint participation in a social relationship.

Some have distinguished between 'backward-looking' and 'forward-looking' notions of responsibility (Goodin, 1995; van de Poel, 2011). Backward-looking notions of responsibility have to do with actions that an agent has already performed, whilst forward-looking notions address the obligations people might have to performing various future actions. My focus here (as well as the focus of a number of views with regards to implicit bias that I will shortly present) is primarily on backwards-looking notions of responsibility: on the question of, given the nature of implicit biases, whether agents could be morally responsible for the implicitly biased actions that they have already performed.

Here are some examples of how philosophers concerned with implicit bias understand the notion of moral responsibility. Jules Holroyd (2015), for instance, maintains that an agent is morally responsible when they are liable to praise or blame, or other sorts of sanctions, and suggests that these are appropriate if the agent ϕ-ed *intentionally*, where ϕ-ing violates some moral standard.

> To say that the agent is blameworthy, then, is to say that they have intentionally done something that violated a moral standard that we expected them to maintain, and as a result certain responses would be warranted: disapprobation or other forms of informal sanction on the part of others; resentment on the part of the wronged party; guilt on the part of the wrong-doer, and resolution to avoid such behaviours or actions in future (indeed, to take responsibility for that) (Holroyd, 2015: 513).

Neil Levy (2014a) also couches moral responsibility in terms of how agents are permissibly treated by others in the wake of acting, suggesting that blameworthiness informs further social interactions, such as the distribution of burdens. Accordingly

> …an agent is morally responsible for an action or omission if the fact that they have performed that action, in the circumstances and manner in which

> they acted, is relevant to how they may permissibly be treated when it comes to the distribution of benefits and burdens. To say an agent is blameworthy for an action, for instance it to say that *(ceteris paribus)*, because they have acted in that way, they may permissibly be punished or that, if burdens are to be distributed, it is better that they fall on them rather than on others who are not blameworthy (Levy, 2014a: 2-3).

As with the notion of 'agential', my aim in this thesis is not to offer an account of moral responsibility. Instead, I will present a number of accounts on which it is argued that implicit biases, and implicitly biased actions, fail to meet some of the preconditions for their agents to be considered morally responsible. I will then respond arguing that at least some implicit biases, and implicitly biased actions *do* meet these conditions which SDR theorists maintain are necessary for moral responsibility.

Let us now turn to the SD and SDR views.

2.2. Substantial Distinction (SD) Accounts

In the following, I give an overview of the substantial distinction (SD) arguments, that there is a distinction in kind between our implicit biases and the actions which flow from them, and attitudes that we attribute to persons and think of as agential (such as beliefs and desires) and the actions which flow from them. I also give an overview of the related substantial distinction: responsibility (SDR) arguments that, in the manner that they differ from beliefs, and belief guided actions, implicitly biased actions fail to meet at least some of the criteria necessary for moral responsibility, and so we can be morally responsible neither for having implicit biases, nor for manifesting implicit bias in our actions. For clarity of exposition, I will label the various SD(R) theorist's main claims with their author's initials. I will then assess these claims in the final section of this chapter and identify three common themes: those of (i) awareness, (ii) structure and processing, and (iii) control.

2.2.1. *Jennifer Saul (2013)*

Jennifer Saul characterises implicit biases as 'unconscious', suggesting that:

> The implicit biases that we are concerned with here are unconscious biases that affect the way we perceive, evaluate, or interact with people from the

> groups that our biases "target". …in the case of women in philosophy, implicit biases will be unconscious biases that affect the way we judge (for instance) the quality of a woman's work, leading us to evaluate it more negatively than it deserves… (Saul: 2013, 40)

She also contrasts implicit biases with 'traditionally understood bias' which she suggests is constituted by *conscious* belief (2013: 39-40).

> By focusing on the phenomena that I discuss in this chapter, I don't mean to suggest that bias as traditionally understood (e.g., the conscious belief that women are bad at philosophy) is a thing of the past. Unfortunately, it does still exist. (Saul: 2013, 39-40)

There are a number of ways in which the categorisation of an entity as conscious or unconscious may be philosophically significant. It might be phenomenologically, metaphysically, or epistemologically significant, for instance. In Saul's case, maintaining that implicit biases are unconscious seems to be an *epistemologically* relevant characterisation, since, in the above quotation, implicit biases are said to distort our epistemic practices—our judgements and evaluations. In other words, for Saul, it would seem that what is significant about our implicit biases, from the agential perspective, is that we do not know about them.

However, as Holroyd (2012, 2015a) has suggested, Saul's characterisation of implicit bias as unconscious is ambiguous. On one interpretation, it is implicit biases, qua *attitudes*, that are unconscious—a person is not conscious that they have the attitude, and is not conscious of its content. On the other, it is the *influence* of a person's implicit biases on their actions that is unconscious. Possibly, Saul thinks that both are the case.

Perhaps owing to the fact that the main focus of the 2013 paper is to make the case for how prevalent implicit bias is likely to be in philosophy departments, and what we ought to do about it, however, Saul's argument is somewhat quick. In the following, oft quoted passage, she makes a number of other claims about what characterises implicit biases, and why we ought not to be held morally responsible for our implicit biases. (Again, this SDR claim is ambiguous between moral responsibility for *having* implicit biases or moral responsibility for the influence of our implicit biases on action):

> A person should not be blamed for an implicit bias of which they are completely unaware that results solely from the fact that they live in a sexist culture. Even once they become aware that they are likely to have implicit biases, they do not instantly become able to control their biases, and so they should not be blamed for them. (They may, however, be blamed if they fail to act properly on the knowledge that they are likely to be biased—e.g., by investigating and implementing remedies to deal with their biases.) (Saul: 2013, 55)

In addition to the claim that we are not consciously aware of our implicit biases, on either of the two interpretations given above, Saul adds a claim about the genesis of our implicit biases: that they result solely from our living in a bigoted culture.[12] This latter claim would seem to be about how we came to have our implicit biases and how much control we have over them: if implicit biases result solely from our culture, then this implies that we do not exert control over the acquisition of our implicit biases. Saul accepts that we are able to become aware that we are likely to have implicit biases, through learning about the empirical studies that reveal the prevalence of such biases and the mitigation strategies that are effective.

As Holroyd has pointed out, the relevant sense in which people are unaware of their implicit biases for Saul is that they cannot *introspect* on their implicit biases or the influence of these biases on action. Holroyd summarises introspective awareness in the following:

> One might have introspective awareness with respect to whether certain beliefs or feelings are playing a role in one's decisions: one can ask oneself, and on reflection give an answer. But, the claim goes, one cannot simply introspect and discern if an implicit bias is operating in the production of action. (Holroyd, 2015: 513)

Indeed, other theorists have claimed that lack of introspective awareness is a distinguishing feature of implicit bias:

---

[12] The quoted claim (and the main focus of Saul's paper) is about implicit biases against women, but we can assume that the point generalises to biases against all stigmatised groups, which Saul does discuss elsewhere in the paper.

> …a wealth of cognitive psychology has demonstrated that we are lousy at introspection. ...implicit biases are those that we carry without awareness or conscious direction. (Kang et al. 2012: 2)

(See also Kelly and Roedder (2008) in the next section for corroboration of the claim as regards introspective awareness.) However, in the second sentence of the last quote from Saul, it is clear that she does concede that people may have some kind of awareness that they have implicit biases, they may have *inferential* awareness of this. After learning about the various data such as that presented in Chapter 1, people can infer that they, like the majority of the population, are likely to be implicit biased. I will investigate the introspective/inferential distinction in much more detail in Chapter 3. Even inferential awareness that one may have implicit biases is not, for Saul, sufficient for being able to exert control over the influence of implicit biases on action. Accordingly, I think that we can attribute the following three key claims to Saul:

JS1: Implicit biases are unconscious (implicit biases and/or their influence on our actions are not available to introspective awareness)

JS2: We are not able to exert control over the acquisition of our implicit biases, in virtue of the fact that they result solely from our living in a bigoted culture.

JS3: Inferential awareness that we are likely to be implicitly biased is not sufficient for control over implicit biases

Saul does not expressly contrast implicit biases with agential states. However, she is committed to an SDR claim—that is, a claim about our responsibility for implicit bias, in virtue of the distinguishing features that she thinks implicit biases have. She suggests that because of JS1, JS2 and JS3 people should not be blamed for their implicit bias: either for having an implicit bias or for acting on its influence.

In order for Saul's blamelessness claims to go through, though, some extra premises are needed:

JS4: It is a necessary condition for moral responsibility for having a mental state *m*/for action influenced by a mental state *m* that the agent is introspectively aware of *m*/that *m* influences action.

JS5: It is a necessary condition for moral responsibility for having a mental state *m*/for action influenced by a mental state *m* that the agent is able to control the acquisition of *m*.

JS6: It is a necessary condition for moral responsibility for action influenced by mental state *m*, that when an agent becomes inferentially aware that she has *m*, she is [instantly] able to control the influence of *m* on action.

It is not clear if Saul thinks that it is the *conjunction* of S1, S2 and S3 (together with the implicitly held S4, S5 and S6) that entail that agents are not morally responsible for implicit biases, or whether each claim is sufficient on its own to deliver blamelessness. I'll explore this in more detail in the chapters to come.

2.2.2. *Daniel Kelly and Erica Roedder (2008)*

Daniel Kelly and Erica Roedder (2008) put forward a case for thinking that certain features distinguish implicit biases from the agential attitudes and actions for which we are morally blameworthy. Like Saul (2013)—or, more precisely, Holroyd's (2015) interpretation of Saul—Kelly and Roedder (2008) think that we lack introspective awareness of our implicit biases:

> Neither introspection nor honest self-report are reliable guides to the presence of such mental states, and one may harbor implicit biases that are diametrically opposed to one's explicitly stated and consciously avowed attitudes. (Kelly and Roedder, 2008: 532)

Their main claims, then, are:

K&R1: People cannot reliably introspect on the presence of their implicit biases

K&R2: The [apparent] content of our implicit biases may be diametrically opposed to the content of our explicitly stated and consciously avowed attitudes

I think K&R2 has to be formulated as a claim about the *content* of the states instead of the states themselves: it not clear what it is for a mental state to oppose another mental state, other than in terms of their contents. But since there is a debate about whether implicit biases are contentful states, at this stage, we should accept that implicit biases might only *seem* to have a content in terms of the way that they affect action, whilst possibly not being the right kind of entities themselves to have a content.[13] (I will explore whether implicit biases have a content, and what sort of content this is, more fully in Chapter 4).

Kelly and Roedder (2008) also consider our reasoning capacities to be 'impaired' in harbouring implicit biases, and say that we have no introspective access to that impairment (2008: 532):

> …if you harbor a racial bias, then you are not responding to reasons in the way that you ought to. (Kelly and Roedder, 2008: 532)

There is some ambiguity in this claim. Firstly, it is not clear whether Kelly and Roedder mean that people do not respond to reasons when they *acquire* an implicit bias, or whether they do not respond to reasons when an implicit bias *influences* action. Secondly, it is not clear whether they mean that implicit biases *could not be* responsive to reasons, or whether we just are not in fact responsive to reasons when we acquire/are influenced by them. So I'll represent all of these possibilities:

> **K&R3:** Implicit biases cannot respond/do not happen to respond to reasons, when acquired/when influencing action.

Another notable characteristic that Kelly and Roedder attribute to implicit biases is evident in the following:

> …the IAT requires subjects to make snap judgments that must be made quickly, and thus without moderating influence of introspection and deliberation and often without conscious intention. Biases revealed by an

[13] For arguments to the effect that implicit biases don't have propositional content, see Gendler, 2008a, 2008b, and Levy, 2014a, and for arguments in favour, see and de Houwer, 2014 and Mandelbaum, forthcoming.

> IAT are often thought to implicate relatively automatic processes. (Kelly and Roedder, 2008: 525)

This would appear to be a claim about how implicit biases operate to influence action. So:

> **K&R4:** Implicit biases influence action automatically

Kelly and Roedder make a claim about the features that establish that we are not morally responsible for harbouring implicit biases in the following passage:

> Particularly in the case of implicit attitudes, it is salient that their acquisition may be rapid, automatic, and uncontrollable. These features, it might be thought, are related to features that establish blameworthiness – such as identification (Frankfurt) or reasons-responsiveness (Fischer and Ravizza). For instance, it might be said that the implicitly racist person doesn't identify with his implicit attitude, or that the attitude isn't responsive to reasons; thus we cannot hold a person fully accountable for those implicit attitudes. If this is right, one might say that such attitudes are morally wrong – and condemnable – but that the person himself cannot be blamed for having them. (Kelly and Roedder, 2008: 532)

Two things should be noted here, before the relevant claims are formulated. Firstly, Kelly and Roedder suggest that they are reluctant to embrace the claim about a lack of responsibility wholeheartedly, should further research render this claim inaccurate: for instance, if narrow-mindedness were to partially explain the acquisition of implicit bias (2008: 532). Secondly, they clarify, in a footnote, the importance of coupling a relevant theory of moral responsibility with the claim that the acquisition of implicit biases is uncontrollable:

> We have stated that it is more *salient* that implicit attitudes are uncontrollable. That's because, arguably, the acquisition of most *explicit* attitudes is uncontrollable as well; it's just not salient at first glance. One does not control one's acquisition of, for instance, one's beliefs about plants, one's attitudes towards pets, etc. So one will need to appeal to more complex or carefully delineated features – perhaps identification or reasons-responsiveness – if one wants to claim that implicit attitudes are not proper

> subjects of blame, but that explicit attitudes are. (Kelly and Roedder, 2008: 537-8fn)

Accordingly, then, we can ascribe to them the following claims:

> **K&R5:** Implicit biases are acquired rapidly, automatically, and uncontrollably
>
> **K&R6:** If moral responsibility for having mental state *m* turns on whether state *m* was not acquired rapidly, automatically, and uncontrollably, then people are not morally responsible for their implicit biases.

Of course, it might be that moral responsibility *does not* turn on whether the relevant mental states are acquired rapidly, automatically, and uncontrollably, but nevertheless *does* turn on something like identification (Frankfurt, 1971) or reasons-responsiveness (Fischer & Ravizza, 1998)—characteristics which turn out to not be present in the case of implicit biases. It is beyond the scope of Kelly and Roedder's (2008) paper to explore this possibility, and so it would be unwarranted to attribute any related claims to them. Since other SD theorists make more direct reasons-responsiveness and control claims (as we shall see shortly) I can focus on those.

2.2.3. *Tamar Szabó Gendler (2008a, 2008b)*

Two things make Tamar Szabó Gendler's position different from the views that we have just seen Saul, and Kelly and Roedder defend. Firstly, Gendler is not concerned with whether we are morally responsible for implicit bias, rather, she is interested in the features which distinguish implicit biases from some agential attitudes—beliefs in particular. Secondly, unlike Saul, and Kelly and Roedder, who discuss features that they think characterise implicit biases, but who do not explicitly state that no agential attitudes *also* share these characteristics, Gendler (2008a, 2008b) does present an explicit case for there being a substantial distinction between implicit biases on the one hand and agential attitudes such as beliefs on the other.

Gendler thinks that implicit biases are "automatic associations" that are "inevitably" encoded, where people typically "disavow the normative content of these associations," (2011: 38-9). For Gendler, implicit biases are a subclass of a

kind of attitude which generates actions that subjects do not seem to fully endorse—a *sui generis* kind of mental state which she calls 'alief'. Aliefs are supposed to be fundamentally different from agential attitudes such as beliefs. Unlike a belief, which might just have a representational component, and which figures in practical reasoning in virtue of its propositional content, a paradigmatic alief is an attitude with a representational component, an affective component and a behavioural component, all of which are "associatively linked" Gendler, 2008a: 642). Gendler spells out this idea further, as follows:

> In paradigmatic cases, activated alief has three sorts of components: (a) the representation of some object or concept or situation or circumstance, perhaps propositionally, perhaps nonpropositionally, perhaps conceptually, perhaps nonconceptually; (b) the experience of some affective or emotional state; (c) the readying of some motor routine. (2008a: 643)

An attitude's being 'associative' means that its constituent concepts are not coupled together in virtue of their propositional contents. This is a technical psychological notion that I will discuss in more detail in Chapter 4, but it raises the possibility that some aliefs may bring about actions which are discordant with an agent's beliefs, in that they contradict the contents of these beliefs.

According to Gendler, aliefs generate actions in a number of instances where subjects have reason to refrain from engaging in such actions. To illustrate the point, let's look at her treatment of another subclass in the alief taxonomy: unwarranted disgust responses. She appeals to a series of experiments from Rozin and colleagues (Rozin *et al.*, 1986; Rozin *et al.*, 1990) which reveal that participants show reluctance to put their mouths on perfectly clean pieces of plastic that are shaped like vomit; prefer not to eat soup from a brand-new bed pan; and generally chose to drink from a glass labelled 'table sugar' over one labelled 'not poison', even though they know there is no poison in either glass. Gendler suggests that whilst they *believe* that the plastic is sterile, for instance, seeing something shaped like vomit activates an alief with the representational content of vomit, as well as the affective content of disgust, which in turn activates an avoidance motor routine.

Unwarranted fear responses are also a part of the alief taxonomy. Gendler notes that some people who believe a high glass platform to be perfectly

structurally sound nevertheless experience feelings of fear and vertigo as they walk across it, with some reluctant to walk across it at all. In this instance, whilst people believe that the bridge is safe, seeing the drop activates an alief with the representational content of height, and the affective content of fear, which, again, activates some avoidance behaviour. Well rehearsed, but misapplied behaviour routines also count as aliefs. In another example, after forgetting her wallet and borrowing cash from a friend, Gendler reaches for her wallet to stow the cash—finding that she does not have it on her person, which was of course the very reason she borrowed the cash in the first place. (In this last example, it is not clear what the affective component of the alief in question is supposed to be, but perhaps this is an example of a non-paradigmatic alief). Note that Gendler is not committed to the idea that an alief's becoming activated entails that a subject *will* carry out a particular behavioural routine, instead she suggests that activation of the alief "renders it more likely that the routine will actually be performed" (Gendler, 2008a: 644).

Whilst Gendler does not give a particularly extended exposition of implicit bias insofar as it is an example of alief, she suggests that it is alief states that bring about cases of implicitly biased action, such as those described in the following:

> An avowed anti-racist exhibits differential startle responses when Caucasian and African faces are flashed before her eyes. (2008b: 553)

She goes on to suggest that:

> What the IAT unquestionably reveals—as its name indicates—are implicit associations...between certain racial categories, and highly-valenced affective content. (2008b: 577)

Whilst the person in the first quotation above believes in racial equality, seeing an 'African' face activates an alief with that representational content, which activates some negative affect, which in turn readies a motor routine for some sort of behaviour—perhaps an avoidance behaviour. And even if the person in question believes in racial equality, and has good reasons for such a belief, according to Gendler, the aliefs which govern her implicit responses do not change. Her aliefs are just not sensitive to the propositional information encoded in her belief,

information which promotes egalitarian actions, not discriminatory actions. But once the representational component of the alief is activated, the (discriminatory) behavioural routine is executed; or, at least, the agent is primed to execute it, and therefore more likely to execute it than she would have been, should the representational and affective components of the alief not have been activated.

Let us take stock, and summarise the key claims so far:

TG1: Implicit biases are aliefs: *sui generis* tripartite mental states with a representational component, an affective component, and a behavioural component, which are 'associatively linked'.

TG2: Activation of the representational content of an implicit bias renders it more likely that an implicitly biased behavioural routine will actually be performed.

Beliefs, according to Gendler, are importantly different to aliefs. She holds that:

> If I believe that *P*, and subsequently learn that not-*P*, I will revise my belief... Learning that not-*P* may well not cause me to cease alieving that *P*... alief just is not reality-sensitive in the way belief is. Its content does not track (one's considered impression of) the world. (2008a: 651)

She also says:

> Beliefs change in response to changes in evidence; aliefs change in response to changes in habit. If new evidence won't cause you to change your behaviour in response to an apparent stimulus, then your reaction is due to alief rather than belief. (2008b: 566)

As Gendler is talking more generally about what she sees as the differences between aliefs and beliefs in these sections, she does not appeal directly to any evidence to support the idea that implicit biases, insofar as they are aliefs, do not update in response to evidence. Perhaps she has in mind the idea that explicit egalitarian attitudes (as measured by self-report questionnaires) appear to co-exist alongside implicit biases (Nosek et al, 2007). If we suppose that people hold their explicit egalitarian attitudes in light of egalitarian reasons, but then observe that they have implicit biases with (apparent) contents that contradict that of the

explicit attitudes, then we might conclude that implicit biases have failed to respond to the reasons that the person in question evidently sees there to be when forming and acting in accordance with their explicit beliefs.[14]

When Gendler suggests, in the above passages, that beliefs *will* be revised, and that they change in response to evidence, whilst aliefs do not, it is not clear whether these are descriptive or normative claims. If descriptive, then the claim is that beliefs in fact *do* change in response to evidence, whilst aliefs, in fact, do not. But then consider what Gendler says elsewhere:

> …belief *aims* to 'track truth' in the sense that belief is subject to immediate revision in the face of changes in our all-things-considered evidence. When we gain new all-things-considered evidence—either as the result of a change in our evidential relation to the world, or as a result of a change in the (wider) world itself—the *norms of belief require* that our beliefs change accordingly. (2008b: 565, emphasis mine)

If normative, then the claim is that beliefs are governed by norms that require that they change in response to evidence, even if, in fact, they do not always change accordingly, whilst aliefs are either governed by no norms at all, or they are governed by norms that differ from those which govern belief. (Although, as I shall argue in Chapter 4, neither possible claim succeeds as a means to distinguish beliefs from aliefs in general, and implicit biases specifically.)

So, we have some further important claims as regards implicit biases (insofar as they are aliefs), and how they supposedly differ from beliefs:

TG3: Implicit biases, insofar as they are aliefs, are not sensitive to the propositional information encoded in mental states such as beliefs: Learning that not-*P* may well not cause me to cease having an implicit bias with the apparent content that *P*.

TG4: Beliefs change in response to changes in evidence, implicit attitudes change in response to changes in habit. If new evidence won't cause you to change your prejudiced behaviour in response

[14] Indeed, Gendler's remarks in the Philosophy TV debate with Schwitzgebel appear to support this interpretation. See http://www.philostv.com/2010/09/02/tamar-gendler-and-eric-schwitzgebel/

> to an apparent stimulus, then your reaction is due to an implicit attitude rather than belief.[15]

2.2.4. *Neil Levy (2013, 2014a, 2014b, 2015)*

Neil Levy has written extensively about implicit attitudes generally, and more specifically on whether implicit biases are agential or non-agential attitudes, as well as on whether agents are morally responsible for implicit bias. I am going to focus mainly on the thorough exposition that Levy gives in his 2014 book, *Consciousness and Moral Responsibility* (denoted as '2014a'), which builds on ideas presented in the 2013 paper, and which, very helpfully, gives clear desiderata for when an attitude is supposedly available to conscious awareness.

Levy's aim in the book is to present a comprehensive SDR account: that is, an account on which we are only morally responsible for actions generated by attitudes of which we have a particular kind of awareness, which will be defined shortly. The ensuing SDR argument may be summarised in what Levy calls 'the consciousness thesis.' In the interests of recording Levy's key claims, as I have been doing for all SD(R) theorists in this chapter, I'll take the whole quotation in which the consciousness thesis is stated, and label it as the first key claim, as follows:

> **NL1:** …only when we are conscious of the facts that give our actions their moral significance are those actions expressive of our identities as practical agents and do we possess the kind of control that is plausibly required for moral responsibility, (2014a: 1).

As Levy argues in the 2014 book, if an attitude fails to be conscious, then it does not constitute the agent's evaluative stance—the perspective from which they act agentially—and, consequently, the agent cannot be morally responsible for any actions guided by the attitude in question. (And so, as we will see shortly, the SDR argument relies on the success of a related SD argument.)

Let us explore Levy's argument in more detail. Levy has a nuanced notion of the sort of conscious awareness an agent needs to have to render the decisions and actions which flow from these attitudes as agential. Like Saul, he is not

---

[15] As mentioned above, this is currently ambiguous between a normative interpretation or a descriptive interpretation. I will discuss this further in Chapter 4.

interested in the phenomenal content of conscious attitudes, but in their informational content, and in whether this is readily accessible for use in reasoning, judgement and action (Levy, 2014a: 29). As Levy acknowledges, this notion of accessibility requires further analysis, for a state may be available to be used in reasoning, say, because the agent in question is *occurrently* aware of its content, or because it has a *dispositional* profile such that the agent would become occurrently aware of it in some possible scenario. Levy suggests that occurrent awareness is too demanding a notion for the sense of accessibility that is relevant to agency, because our *beliefs* may guide agential behaviour, even though we are not consciously attending to them at every moment during this process (2014a: 31). Consider driving home 'on autopilot', for instance, and indicating to turn left without attending to the fact that you have done so—it would seem incorrect to suggest that because you were not occurrently aware that the next left is the way to your house, your action was therefore not agential. As such, for Levy, occurrent awareness is too restrictive a notion of the kind of consciousness that is necessary for agency and moral responsibility.

The above suggests that the conscious awareness necessary for agency is dispositional in kind. Accordingly, one may count as conscious that *P* in the actual sequence of events, as long as there is some counterfactual scenario in which one would become occurrently aware that *P*. However, one might think that there is a lower boundary on the closeness of counterfactual scenarios in which one would become occurrently aware that *P*, below which it no longer makes sense to consider one as conscious that *P* in the actual sequence of events. To see this, suppose that I am not occurrently aware that today is your birthday. Further, suppose that the only counterfactual scenario in which I would become occurrently aware that today is your birthday, is if I were to see two red balloons floating in the sky. If this is the only scenario in which I would recall the information that it is your birthday, then one might think that it is odd to suggest that in the actual sequence of events, where it is very unlikely that I'll see two red balloons, I was nonetheless consciously aware that it is your birthday. For Levy, I should not be held morally responsible for forgetting that it is your birthday in this sort of case (Levy, 2014a: 34). As such, Levy thinks that unrestricted dispositional awareness is too permissive a notion of the kind of consciousness that is necessary for agency and moral responsibility.

For Levy, this does not necessitate that we revert back to occurrent awareness to characterise the kind of consciousness that is necessary for agency and moral responsibility (which, as we saw above, was too restrictive), but that we apply some restrictions to the dispositional account. In light of this, Levy proposes a new understanding of awareness that he terms 'personally availability' to characterise the kind of awareness that is necessary for agency and moral responsibility. An attitude is personally available if it fulfils two conditions: (i) being online; and (ii) being effortlessly recallable, (2014a: 33). An attitude is online if it is currently guiding some behaviour and it is effortlessly recallable if it *would* become occurrently conscious in the presence of a large range of 'ordinary cues' (2014a: 34):

> Information is available for easy and effortless recall if it would be recalled given a large range of ordinary cues: no special prompting (like asking a leading question) is required. For instance, for [an agent who is not occurrently aware that it is her friend's birthday] to have the information personally available to her, the presence of a telephone would likely cause her to be occurrently aware of her friend's birthday. (2014a, 34)

From now on, I will notate Levy's specialised notion of consciousness as personal availability, as it is defined above, as 'consciousness$_{PA}$'. Accordingly we get:

> **NL2:** A state is conscious$_{PA}$ when it is (a) online, and (b) effortlessly recallable. Being conscious$_{PA}$ is a necessary condition for mental states and their guidance of actions to be agential.

Levy does not detail a set of conditions that makes an item or an event an 'ordinary cue' for the effortless recall of an attitude, and so this might be up to our intuitions on a case by case basis. In his defence, we might think that there is a sense in which a telephone is more likely to prompt effortless recall of the information that it is a friend's birthday than, say, a desk or a window, as well as a sense in which a telephone is a more 'ordinary' cue for recall than two floating balloons of a specific colour. I do, however, think that this notion is far from watertight, and I am not sure that it will do the required work, as I will argue in Chapter 3.

In later chapters of his 2014 book, Levy argues that consciousness$_{PA}$ matters for agential and morally responsible actions, because an agent's attitudes being *integrated* with each other is a necessary feature of a set of agential attitudes.[16] For Levy, an agent's attitudes are integrated if they have broadly consistent contents: when an agent detects that they have attitudes with content that contradict the content of other attitudes, these are either updated or rejected. Levy argues that attitudes may only be integrated with each other in accordance with their content if they are conscious$_{PA}$, and maintains that this claim is supported by various evidence. For instance, according to Baumeister and Masicampo (2010), when nonconscious systems are primed with two-word phrases, each word has an independent priming effect, which suggests that some unconscious processes are not sensitive to semantic content in the way that some conscious processes are, and therefore that they are unable to integrate attitudes in accordance with their semantic content. Levy also appeals to evidence from Deutsch, Gawronski and Strack (2006); Hasson and Glucksberg (2006) which suggests that some nonconscious processes are blind to other logical constructs such as negation (or, at least, they tend to represent the contents of a negated term as asserted).[17] Summarising the relevant results, Levy says:

> Activating concepts nonconsciously has effects on subjects' attitudes, but these effects are associative and not logical. All of this appears to be evidence of an absence of the capacity to integrate the content of representations; whereas nonconscious processing of contents may cause the activation of semantically related content, only when the processing is conscious is the activation logically coherent. Priming contents facilitates access to semantically related contents, but not in a coherent or integrated manner. (2014a, 53)

For Levy, these results show that when a person's attitudes are activated and processed nonconsciously, their logical contents are not preserved or integrated with the person's other attitudes. The SD arguments of both Gendler and Levy rely on there being a fundamental difference between 'associative' and 'logical'

---

[16] See chapters 3 and 4 of Levy's 2014a for a full exposition of this view.

[17] Actually, the research that Levy cites in fact suggests that such processes may be *trained* to represent negation, but without such training they remain typically bad at doing so.

processing, and both rely on results in the empirical literature to back up this supposed distinction. I will give a more detailed and critical exposition of the relevant empirical claims which allegedly support the distinction between 'associative' and 'logical' in Chapter 4. For present purposes, it is sufficient to say that Levy's claim is that when informational states are activated and processed nonconsciously, they cannot be integrated with the agent's "personal level" attitudes, by which I think he means something close to 'agential' attitudes as discussed in §2.1.1.

Levy motivates the idea that integration of states is a necessary feature of a set of personal-level attitudes by making reference to somnambulists, who perform complex behaviours but are in a sleep-like, and so nonconscious, state at the time. He suggests that there is no basis on which we can attribute somnambulist actions to agents, for the very reason that the attitudes that guide somnambulist actions are not integrated with personal level attitudes. One of the starkest examples in the legal responsibility literature is the case of Ken Parks, who attacked his step-parents with a kitchen knife, presented in Broughton *et al.* (1994).[18] Parks was acquitted on the grounds that he acted during a state of somnambulistic automatism, a result which Levy argues is philosophically significant because it shows that when the information which guides behaviour is not properly integrated with the rest of an agent's attitudes, we do not think that the resulting behaviour reflects their agency, and so this behaviour is not the kind of thing for which an agent may be morally, or legally, responsible.

Let us summarise Levy's relevant SD claim as follows:

NL3: Being integrated with each other is a necessary condition for mental states and their guidance of actions to be agential.

---

[18] The case occurred in May 1987, when Parks, a Canadian man, rose from his bed one night and climbed into his car, and driving several blocks to the home of his parents-in-law, he retrieved a knife from the kitchen and proceeded to stab both in-laws repeatedly, causing the death of one of them. There was enough evidence to suggest that he carried out his actions during a state of somnambulistic automatism. According to the details of the case, Parks did not have any particular ill-will against his in-laws, at least not that he was $\text{conscious}_{\text{PA}}$ of, and believed that stabbing someone to death is wrong. According to Levy, because the attitudes which guided Parks' somnambulism were neither online, nor effortlessly retrievable, crucially, Parks was unable to compare his actions for consistency with his personal level attitudes. His actions and the attitudes which guide them remained unintegrated with his personal level attitudes.

Let us now turn to Levy's position on implicit bias. Whilst implicit biases might meet the first condition of Levy's notion of 'conscious$_{PA}$', in that they can be online and guiding behaviour, they fail to meet the second condition, because we are apparently not able to effortlessly recall—or even to recall at all—the contents of our implicit biases (Levy, 2014a: 95).[19] So we can ascribe to Levy:

NL4: Implicit bias is not conscious$_{PA}$.

Levy supports this claim with reference to a range of empirical results, many of which I discussed in Chapter 1, and in particular with reference to the Uhlmann and Cohen (2005) police chief hiring experiment, arguing that participants "lacked the ability to detect the processes that generated their confabulated criteria" and "lacked the capacity to see that the choice was not in fact objective" (2014a: 95).

In addition to the claim that implicit attitudes are not conscious in the sense required for them to be personally available, Levy (at least, in his 2014 book and 2014 paper) agrees with Gendler (2008a, 2008b) that implicit attitudes are associatively structured:

> …though implicit attitudes may have quite broad contents, these contents are bound together associatively rather than propositionally. (Levy, 2014b: 31)[20]

Levy further suggests that implicit attitudes are encoded following the repeated co-occurrence of two stimuli, suggesting that implicit biases are:

> …probably acquired by associative systems which respond to regularities in the environment. (2014a: 98)

Accordingly:

---

[19] I will argue against this claim in Chapter 3.

[20] Levy (2015) revises this claim in response to a paper from Mandelbaum (forthcoming), which presents a range of evidence to suggest that implicit attitudes encode, and update in accordance with propositional information. In the 2015 paper, Levy argues that implicit biases are not quite like beliefs, but not quite like straightforward associations either. I will discuss both Mandelbaum's argument that implicit attitudes are propositional (forthcoming) and Levy's (2015) response in Chapter 4.

NL5: Implicit biases are associative, not propositional, in structure.

Levy echoes arguments made by Gendler (2008a, 2008b), to demonstrate that implicit biases are not judgement-dependent, and that it is misleading to think of them as beliefs.

> ...there is good reason to think that the claim that implicit attitudes belong to the class of judgement-dependent attitudes carves up the territory in such a way as to obscure central characteristics of such attitudes. These attitudes are, as we just saw, acquired in ways that bypass rational control, and they are altered in ways that resemble those in which they are acquired. Indeed, as Gendler (2008[b]) suggests, insensitivity to reasons is what distinguishes 'aliefs' (a category of mental state that overlaps considerably with implicit attitudes), from beliefs... Implicit attitudes are *not* judgement-dependent. It is misleading to regard them as a subcategory of the state 'belief'; misleading because it masks the fact that judgement-insensitivity is the hallmark of such states. It is characteristic and perhaps even definitive of such states that they do not respond to our reasons... (2014:99)

Accordingly:

NL6: It is characteristic and perhaps even definitive of implicit biases that they do not respond to our reasons.

Further, for Levy, the associative nature of implicit attitudes, and their resistance to reasons, has implications for the kind of control that we are able to exercise over our implicit attitudes, once we know about them. He suggests:

> we can influence our implicit attitudes only indirectly: by the same kinds of methods whereby we acquired them in the first place (by attempting to form new associations). Whether these methods are arduous, slow, and extremely uncertain (Devine, 1989), or on the contrary relatively rapid, remains controversial (Dasgupta, 2013). (Levy, 2014a: 99-100).

Accordingly:

NL7: We can influence our implicit attitudes only indirectly, by attempting to form new associations.

The notion that implicit biases do not respond to propositional information encoded in our reasons plays an important role in Levy's SDR argument. He suggests that implicit biases "express nothing more than facts like: there is a statistical association between being male and being a police chief" (2014: 102) and that:

> In expressing these attitudes, we do not express anything that is a target of moral condemnation: the fact that I associate *X* and *Y*, nonconsciously, is no basis for holding me morally responsible. (2014: 102)

Finally, then:

NL8: The fact that I associate *X* and *Y*, nonconsciously, is no basis for holding me morally responsible.

Having established that implicit biases are not $\text{conscious}_{\text{PA}}$, and that they are not appropriately structured to respond to reasons, Levy then suggests that we cannot be morally responsible for actions influenced by implicit bias:

> When agents are aware neither of the mental states that are responsible for the moral significance of an action, nor of that moral significance in itself, neither states nor significance is globally broadcast, and the agent cannot assess either for consistency or conflict with their personal level beliefs. The action therefore does not express their evaluative agency. There are good reasons to think that actions like this are not even expressions of morally significant implicit attitudes that cause them. The attitudes involved do not have the right kind of contents to play the role of reasons for actions: they are too disunified and too thin for that, and they are neither acquired nor maintained in a manner that is regarded by the agents (even nonconsciously) as reason giving. Insofar as moral responsibility depends on expression (of evaluative agency or even of attitudes), we ought to deny that agents are morally responsible for these actions. (2014a: 102-103).

2.3. ORGANISING THE KEY ARGUMENTATIVE CLAIMS

As suggested earlier, I think that three themes emerge from the SD and SDR arguments summarised in the foregoing section. Broadly speaking, they are:

> (I) AWARENESS: we lack awareness of our implicit biases, and lack awareness that they influence actions
> (II) STRUCTURE & PROCESSING: implicit biases are associative (rather than propositional), and so are not structured in a manner appropriate to enter into logical inferences
> (III) CONTROL: we lack control over the formation or modulation of our implicit biases, and over the influence of our implicit biases on our actions.

Clearly, not all of the arguments presented in §2.2 support just one claim out of (I)-(III). Further, the themes are somewhat interrelated, and some arguments are mutually supportive. For example, control claims are sometimes based on either awareness, or structure and processing claims (or both). To reflect this, my discussion of each theme in the chapters to come will take into account arguments from the other themes. However, I think that categorising the key arguments from different philosophers in accordance with the three categories as above is useful because there are clearly some common themes, which it makes dialectical sense to address concurrently. Accordingly, each of the three chapters to follow focuses on one of the three themes.

The claims by the various philosophers discussed are classified under the three headings as follows:

(I) AWARENESS

SD claims based on awareness

JS1: Implicit biases are unconscious (implicit biases and/or their influence on our actions are not available to introspective awareness)

K&R1: People cannot reliably introspect on the presence of their implicit biases

NL2: A state is conscious$_{PA}$ when it is (a) online, and (b) effortlessly recallable. Being conscious$_{PA}$ is a necessary condition for mental states and their guidance of actions to be agential.

NL4: Implicit bias is not conscious$_{PA}$.

SDR claims based on awareness (I)

JS4: It is a necessary condition for moral responsibility for having a mental state *m*/for action influenced by a mental state *m* that the agent is introspectively aware of *m*/that *m* influences action.

NL1: …only when we are conscious of the facts that give our actions their moral significance are those actions expressive of our identities as practical agents and do we possess the kind of control that is plausibly required for moral responsibility, (2014a: 1). (*Also in control section)*

(II) STRUCTURE AND PROCESSING

SD claims based on structure and processing

K&R2: The [apparent] content of our implicit biases may be diametrically opposed to the content of our explicitly stated and consciously avowed attitudes

K&R3: Implicit biases cannot respond/do not happen to respond to reasons, when acquired/when influencing action.

TG1: Implicit biases are aliefs: *sui generis* tripartite mental states with a representational component, an affective component, and a behavioural component, which are 'associatively linked'.

TG3: Implicit biases, insofar as they are aliefs, are not sensitive to the propositional information encoded in mental states such as beliefs: Learning that not-*P* may well not cause me to cease having an implicit bias with the apparent content that *P*.

TG4: Beliefs change in response to changes in evidence, implicit biases, insofar as they are aliefs, change in response to changes in habit. If new evidence won't cause you to change your prejudiced behaviour in response to an apparent stimulus, then your reaction is due to an implicit attitude, insofar as it is an alief, rather than belief.

NL3: Being integrated with each other is a necessary condition for mental states and their guidance of actions to be agential.

NL5: Implicit biases are associative, not propositional, in structure.

NL6: It is characteristic and perhaps even definitive of implicit biases that they do not respond to our reasons.

SDR claims based on structure and processing

NL8: The fact that I associate *X* and *Y*, nonconsciously, is no basis for holding me morally responsible.

(III) CONTROL

SD claims based on control

JS2: We are not able to exert control over the acquisition of our implicit biases, in virtue of the fact that they result solely from our living in a bigoted culture.

JS3: Inferential awareness that we are likely to be implicitly biased is not sufficient for control over implicit biases

TG2: Activation of the representational content of an implicit bias renders it more likely that an implicitly biased behavioural routine will actually be performed.

K&R4: Implicit biases influence action automatically.

K&R5: Implicit biases are acquired rapidly, automatically, and uncontrollably

NL7: We can influence our implicit attitudes only indirectly, by attempting to form new associations.

SDR claims based on control

JS5: It is a necessary condition for moral responsibility for having a mental state *m*/for action influenced by a mental state *m* that the agent is able to control the acquisition of *m.*

JS6: It is a necessary condition for moral responsibility for action influenced by mental state *m*, that when an agent becomes

inferentially aware that she has *m*, she is instantly able to control the influence of *m* on action.

K&R6: If moral responsibility for having mental state m turns on whether state m was not acquired rapidly, automatically, and uncontrollably, then people are not morally responsible for their implicit biases.

NL1: …only when we are conscious of the facts that give our actions their moral significance are those actions expressive of our identities as practical agents and do we possess the kind of control that is plausibly required for moral responsibility, (2014a: 1). *(Also in awareness section)*

SUMMARY OF CHAPTER AND DIALECTIC TO COME

In this chapter, I have presented a number of SD views, according to which there is a substantial distinction between a) our implicit biases and the actions which they influence, and b) attitudes that we attribute to persons insofar as they are agential (such as beliefs and desires) and the actions which they guide. I then outlined the arguments of a subset of SD theorists (the SDR theorists) that, in the manner that they differ from agential attitudes and actions, implicit biases and implicitly biased actions fail to meet at least some of the criteria necessary for moral responsibility.

In response to the above arguments, some philosophers, such as Natalia Washington and Daniel Kelly (2016), concede that implicit biases may well be distinct from agential attitudes, but reject the SDR arguments which proceed from the relevant SD claims—arguing instead that the actions for which we may be morally responsible extend beyond that set of actions that are guided by agential states, and that we have resources at our disposal to stop them manifesting in action. Washington and Kelly defend an account on which moral responsibility for action tracks two things. Firstly, it tracks the role that a person plays in the fair distribution of social goods; and, secondly, it tracks our expectations that, depending on their role, a person *should* know about the empirical findings on implicit bias and the relevant mitigation strategies, rather than tracking whether or not the action in question was guided by agential attitudes. On Washington and Kelly's account, whilst implicitly biased actions are not guided by agential attitudes, a person may still be morally responsible for them if they (i) occupy a role in which they are expected to distribute goods fairly (such as a person who

regularly makes recruitment decisions), and (ii) *fail* to put in place measures that prevent their implicit biases from manifesting (such as anonymising C.V.s).

I will not pursue this sort of response to the SDR arguments.[21] Rather, I shall argue that, for each of the features that SD theorists propose that agential attitudes have, and that implicit biases lack (features on the basis of awareness, processing and control) there is, in fact, *no principled way to maintain a substantial distinction* between (i) all implicit biases, and the actions that they influence, and (ii) all agential attitudes, and the actions that they guide. That is to say that, even when considering the main SD arguments in the philosophical literature on implicit bias, there is in fact no single characteristic that all beliefs, and belief-guided actions have, that all implicit biases, and implicitly biased actions lack (and *vice versa*). As such, I will argue that the SD view does not reflect reality, and should be rejected.

In light of the failure of the SD account, I will be defending what I call the 'continuum thesis', on which implicit biases and beliefs are not discontinuous from one another, and may be ordered on a continuum in accordance with the level of awareness and control that we have of them. One extreme end of this continuum may be populated only by implicit biases (and related actions), while the other extreme end may be populated only by beliefs (and related actions). However, in the middle, there is a considerable area of overlap in which we find both a significant number of implicit biases (and implicitly biased actions) as well as many beliefs, (and belief-guided actions).

Because SDR arguments rely on the truth of SD arguments, by showing that there is no substantial distinction between implicit biases and implicitly biased actions, and the more familiar agential cognitions such as beliefs, and belief-guided actions, we get a refutation of the SDR argument 'for free': Insofar as some implicit biases, and the actions which flow from them, share characteristics with some agential attitudes and actions—namely those in the overlap zone of the continuum—if it is appropriate to hold agents morally responsible for the latter agential attitudes or actions, then it is also appropriate to hold agents morally responsible for the former implicitly biased attitudes or

---

[21] I am convinced by Holroyd's (2015: 517) response to Washington and Kelly's (2016) argument, that almost everyone will turn out to be a goods distributor in a variety of social interactions, but it is unreasonable to suggest that, therefore, everyone ought to know about the relevant empirical findings on implicit bias, as I outline in more detail in Chapter 6.

actions. I will show that it is indeed sometimes appropriate to hold agents morally responsible for agential actions in the overlap zone, and so it is also appropriate to hold agents morally responsible for the implicitly biased actions which share features with these former agential actions, as I will demonstrate in Chapter 6. I will argue that trying to save the SD(R) account by insisting that those beliefs in the overlap zone are not agential after all, considerably restricts the account of human agency, because we end up having to accept that a significant set of human activities, some of which epitomise human flourishing, are not agential after all—and this commits us to an unsatisfactory and incomplete picture of human agency.

With this map of the dialectic in place, let us now turn to the first set of SD claims: those made on the basis of awareness.

CHAPTER 3: RESPONDING TO SUBSTANTIAL DISTINCTION CLAIMS ON THE BASIS OF AWARENESS

In the previous chapter, we saw that a number of philosophers think that there is a fundamental distinction in kind between (i) implicit biases, and the actions that they guide; and (ii) agential attitudes such as beliefs, and the actions that they guide, on the basis of characteristics had by (ii) that are not had by (i) (and *vice versa*). I called this the 'substantial distinction' (SD) view. In this chapter, I shall examine and respond to arguments for the SD view of implicit bias on the basis of the agent's awareness. The central claim is that none of the arguments succeed: no substantial distinction between implicit biases and beliefs (and the influence of each on action) can be established on the basis of the kind of awareness that we have of each. Rather, I will argue that we have the same kind of awareness of at least some of our implicit biases, and their influence on our actions, as we have of at least some of our beliefs, and their guidance of our actions. The SD view cannot accommodate this data. I argue, instead, that with respect to our awareness, implicit biases and beliefs lie on a continuum, on which there is significant overlap between the set of implicit biases and the set of beliefs.

To show this, I outline the three senses of awareness which, according to Holroyd (2015), have been at issue in the philosophical literature on implicit bias: (a) introspective awareness; (b) inferential awareness; and (c) observational awareness, (§3.1). I examine whether there is a substantial distinction between implicit biases, and their influence on actions, and beliefs, and their influence on actions, on the basis of each of these senses of awareness (§3.2). Following Holroyd (2015), I demonstrate that we have as much inferential and observational awareness of at least some of our implicit biases, and of their influence on our actions, as we do of at least some of our beliefs, and of their guidance of our actions. There are various accounts of introspective awareness in the philosophical literature, and so whether agents have introspective awareness of their implicit biases depends on which account one adopts. I argue that if we adopt Borgoni's (2015) 'ordinary' notion of introspective awareness, on which we may use observations of our own mental states as evidence of what we believe, then we may count as introspectively aware of at least some of our implicit biases

and their influence on our actions. This supports the continuum thesis, on which there no substantial distinction between implicit biases and beliefs (and their influence on action) on the basis of introspective awareness.

Holroyd maintains that it is not possible to be introspectively aware of the influence of implicit bias on action, and so she perhaps is working with an account of introspection different from Borgoni's. However, Holroyd also maintains that there are some cases of arguably agential actions in which people are not introspectively aware of the attitudes which guide such actions. This would seem to count against the SD theory. But Holroyd's argument is open to an objection articulated by Levy in his 2014 book. As we saw in the previous chapter, Levy argues that agents do not need to be *occurrently* introspectively aware that a particular attitude guides action in order to count as having introspective awareness of the attitude's guiding role. For Levy, as long as an agent would effortlessly recall the attitude's role in action in the presence of an 'ordinary cue' for that attitude,[22] they count as introspectively aware that the attitude in question guides action. So, SD theorists may argue that agents in the sorts of situations described by Holroyd, whilst not occurrently introspectively aware that their beliefs guide their actions, would nonetheless effortlessly recall these facts in the presence of appropriate ordinary cues, and would, therefore count as introspectively aware that their beliefs guide their actions. If this argument succeeds, then there will be a substantial distinction between our awareness of how implicit biases guide actions as compared to beliefs—namely that we may only effortless recall the influence of the latter, and not the former, in the presence of an ordinary cue.

In response to this argument, I show that it is not the case for all belief-guided actions that, in the presence of an ordinary cue for the relevant belief, the agent will recall the guiding role of that belief—however, effortless recall is not a necessary condition for an action to be agential. As a result, even those who do not rely on Borgoni's ordinary notion of introspection, will not be able to identify a substantial distinction between our introspective awareness of how our implicit biases and beliefs guide actions. Even though we may fail to effortlessly recall the

---

[22] As we saw in Chapter 2, Levy relies on an intuitive notion of what sort of objects count as ordinary cues for particular mental states. For instance, a telephone counts as an ordinary cue for the notion that it is a friend's birthday, (2014a: 34). I will explore this account more critically in this chapter.

influence of the former, we may also sometimes fail to effortlessly recall the guidance of the latter, but this alone does not render the actions in question as non-agential: such actions may well fulfil personal level goals and objectives.

In the final section of the chapter (§3.3) I give a positive account of the awareness that we have of at least some of our implicit biases. There, I introduce the notion of an 'observable class preference'. An agent has an observable class preference when they (1) have made multiple evaluations of some objects of which they are introspectively aware, and (2) the objects in question belong to the same class and are evaluated to have qualities of the same kind and valence. I give four examples of everyday observable class preferences, to demonstrate that they are a sufficiently common phenomenon in much of our everyday agential action. I then argue that at least some implicit biases are observable class preferences. This argument works regardless of which account of introspective awareness described above we endorse. If one accepts Borgoni's account of introspective awareness, on which making inferences from evidence of our own psychology counts as introspecting, then it is plausible to argue that we can discover our observable class preferences through acts of introspective awareness. If one accepts an account of introspective awareness other than Borgoni's, on which we *don't* count as introspecting if we make inferences from any kind of evidence, including the psychological, then it is plausible to argue that we discover our observable class preferences through acts of observational awareness. Regardless of whether one holds Borgoni's (2015) account or not, we have the same kind of awareness of those implicit biases which are observable class preferences as we do of at least some of our everyday observable class preferences. The SD theory is inconsistent with this result. Instead, we must adopt a continuum thesis to account for the case of observable class preferences. I defend this account against three objections.

In light of my reply to SD arguments on the basis of awareness above, I then consider the status of the related SDR claim that our (apparent) lack of awareness renders us unable to be morally responsible for actions influenced by implicit bias. As noted in the previous chapter, SDR theorists rely on arguments about awareness of the guiding role of implicit bias to show that we are not morally responsible for actions influenced by implicit bias. My response is that, if it turns out that we do lack moral responsibility for our implicit biases and their influence on our actions, it will not be because we lack awareness of them, but

perhaps because of some other distinguishing features, features that I will discuss in more detail in the following chapters.

3.1. Three Kinds of Awareness

As we saw in Chapter 1, the idea that implicit biases are necessarily unconscious states, states of which participants could never become aware, was not part of the original distinction between the implicit and the explicit: the original distinction was made on the basis of controlled *vs.* automatic process (Payne and Gawronski, 2010). As we saw in Chapter 2, however, many philosophers argue that the data reveals that we lack awareness of our implicit biases and their influence on our actions. In a 2015 paper, Holroyd demonstrates that there are three senses of awareness in play in philosophical discussions of implicit bias, and puts forward the case for distinguishing them. They are:

- introspective awareness
- inferential awareness
- observational awareness

Holroyd's 2015 paper is about the epistemic preconditions for moral responsibility, and whether we meet these conditions in the case of actions influenced by implicit bias. She is interested in which kind of awareness of attitudes and their influence on action, if any, is necessary to have in order to be morally responsible for the action in question. Ultimately, Holroyd argues that only one of the three kinds of awareness (observational awareness) is a plausible epistemic condition for moral responsibility. She maintains that we do in fact meet this condition in the case of implicitly biased actions, and that if any conditions absolve us of moral responsibility for implicitly biased actions, then they will not be conditions based on awareness.

My main focus in this section is on Holroyd's claims regarding the three distinct kinds of awareness that we may have of attitudes and their influence on action. I will briefly outline Holroyd's definitions of each different notion of awareness, before considering, in the following subsection, whether any may be used to uphold the substantial distinction view.

Introspective awareness

It is plausible that we typically think about introspective awareness when questioning whether we have any awareness of implicit bias. However, the philosophical literature on exactly what introspection amounts to is vast, and contains a number of distinct accounts. Some philosophers propose that introspection is a kind of 'inner sense' which enables us to gain awareness of our mental states in a similar way to that in which our other senses deliver information about the external world (Kant, 1781/2009; Armstrong, 1968/1994). Others have argued that introspection is a kind of direct acquaintance with our mental states, in which the introspecting person cannot be wrong that they are entertaining particular mental states with a particular content (Russell, 1912; Shoemaker, 1968; Davidson, 1984). More recently, Borgoni (2015) has argued for what she calls an 'ordinary' account according to which introspecting on one's mental states involves simply noticing and analysing the occurrence of particular psychological events. On this account, introspection is:

> …active self-reflection and analysis of the manifestations of one's mental states, such as thoughts, feelings, memories of one's actions in particular circumstances and other mental states related to the one in question. …by introspecting, one is engaging in an activity in which one directs one's attention to inner and outer manifestations of one's mental states and eventually comes to understand them better. (Borgoni, 2015: 216)

An agent introspects in this way not through some special perceptual faculty, or through some notion of privileged access, but simply by noticing the occurrence of particular psychological events.

Holroyd suggests that agents gain introspective awareness "simply by reflecting on one's internal states and processes," (2015: 513). This much, at least, seems to be consistent with Borgoni's (2015) 'ordinary' account of introspection, on which introspecting agents notice and reflect on the occurrence of psychological events. To have introspective awareness of an implicit association, Holroyd suggests that we would have to have "awareness of the implicit association itself, or its operation" (2015: 514). And she maintains that we do not have this sort of awareness in the case of implicit biases. In particular, she says:

> One might have introspective awareness with respect to whether certain beliefs or feelings are playing a role in one's decisions: one can ask oneself, and on reflection give an answer. But, the claim goes, one cannot simply introspect and discern if an implicit bias is operating in the production of action (2015: 513).

Indeed, as we saw in Chapter 2, this is the sense of awareness that is in play in Saul (2013) and Kelly and Roedder's (2008) arguments. Their key claims were paraphrased as follows:

JS1: Implicit biases are unconscious (implicit biases and/or their influence on our actions are not available to introspective awareness)

K&R1: People cannot reliably introspect on the presence of their implicit biases

I also think that Levy has in mind introspective awareness in his (2014a) claim about consciousness as personal availability, and in the claim that implicit bias is not personally available, which I paraphrased to:

NL2: A state is $\text{conscious}_{\text{PA}}$ when it is (a) online, and (b) effortlessly recallable. Being $\text{conscious}_{\text{PA}}$ is a necessary condition for mental states and their guidance of actions to be agential.

NL3: Implicit bias is not $\text{conscious}_{\text{PA}}$.

As I noted in the previous chapter, Levy's account is more complex because what agents are $\text{conscious}_{\text{PA}}$ of depends in part on what they are disposed to become *occurrently* conscious of when they encounter certain features of their environment. In particular, agents are only $\text{conscious}_{\text{PA}}$ of the states which guide their actions if they would effortlessly recall them in the presence of an ordinary cue for that mental state. Nonetheless, when agents effortlessly recall the guiding role of a mental state, this is an act of introspective awareness. Accordingly, Levy's account of $\text{consciousness}_{\text{PA}}$ also relies on the notion of introspective awareness, with some dispositional conditions. I will give a critical exposition of Levy's (2014a) account in the next section, §3.2.

So, to conclude this subsection, the notion of introspective awareness is in play in a number of the SD claims, and there are number of different accounts in the philosophical literature of exactly how agents introspect on their mental states. These distinctions will become relevant in my critical discussion in §3.2.

Inferential awareness

For Holroyd, inferential awareness is "awareness of the body of knowledge about people's tendencies to harbour, and display, implicit bias" (2015: 514). The object of which one is inferentially aware with respect to implicit bias, then, is that one is very likely to be biased, given the relevant findings in experimental cognitive science (Holroyd, 2015: 513). Accordingly, it is worth highlighting that Holroyd is using the term 'inferential awareness' in a somewhat specialised sense, as compared with other notable accounts in the philosophical literature on the inferential/introspective divide.[23] For Holroyd, a person becomes inferentially aware that they are likely to be implicitly biased by (i) becoming aware of the body of empirical findings which show that the majority of people are implicitly biased, and (ii) inferring that they are likely to be implicitly biased (2015: 514).

Holroyd notes that it is inferential awareness that is at issue in Saul's claim that "Even once [people]…become aware that they are likely to have implicit biases, they do not instantly become able to control their biases" (Saul, 2013: 55; in Holroyd, 2015: 513) relies on the notion of inferential awareness.[24] Inferential awareness is also at issue in Washington and Kelly's (2016) paper (the main thrust of which I outlined briefly in the previous chapter) in which an externalist epistemic condition for moral responsibility is proposed. That is, they maintain that the extent to which people should be expected to know about empirical findings of implicit bias is indexed to the kind of role they play in society, i.e. whether or not their job involves regularly hiring candidates, for instance. I think that Washington and Kelly's (2016) proposal is an interesting one, but I also think

---

[23] In particular, we should not confuse Holroyd's characterisation of the notion of inferential awareness, which is specifically to do with awareness of *empirical findings*, with, for instance, Ryle's (1949) more general notion of inferential awareness of attitudes, which incorporates awareness of empirical findings about attitudes, *as well* as awareness of one's attitudes gleaned on the basis of one's behaviour. Awareness of one's attitudes on the basis of one's behaviour is accounted for in Holroyd's characterisation of 'observational awareness', which I come to in the main text shortly.

[24] This is the claim that I paraphrased to JS3 in Chapter 2—the claim that having inferential awareness that we are likely to be implicitly biased is not sufficient for control over implicit biases.

that Holroyd (2015) raises some difficult problems for their account. I will talk in more detail about Washington and Kelly's (2016) argument, and Holroyd's (2015) critique in Chapter 6.

Observational awareness

The third kind of awareness identified by Holroyd is observational awareness. According to her, this is awareness of the *manifestation* of bias in behaviour (2015: 514). For example, a person undertaking an IAT test might become aware that are matching stereotype-incongruent items more slowly than the speed at which they matched stereotype-congruent items, as discovered by Monteith *et al.* (2001). The claim that people can have observational awareness that they are implicitly biased is supported by a number of empirical findings that I discuss in more detail below.

3.2. Is There A Substantial Distinction, On The Basis Of Any Of These Kinds Of Awareness, Between Implicit Biases and Beliefs?

As mentioned above, Holroyd's aim in her 2015 paper is to discern which, if any, of these kinds of awareness is necessary for moral responsibility. My aim in this section is slightly different. I am interested in whether any of these kinds of awareness may be used to uphold a substantial distinction between implicit biases, and the actions that they influence; and beliefs, and the actions that they guide. I will assess this question starting with the last notion discussed, namely observational awareness.

3.2.1. *The SD thesis and observational awareness*

It seems likely that we sometimes discover what we believe through acts of observational awareness. For instance, I might observe myself being unpleasant to a new acquaintance at a party, and realise that I have taken a disliking to him, which has surfaced in my short comments about his music tastes. I might observe the mess in the kitchen and realise that it didn't bother me enough to tidy it up last time I was in the kitchen. I might observe that I tend to pick up and coo over my tortoiseshell cat more than my tabby cat, and realise that I prefer my tortoiseshell to my tabby. These three examples demonstrate that it is possible to have

observational awareness of at least some of our beliefs and their influence on our actions.

What about observational awareness of our implicit biases? Holroyd (2015) summarises some compelling empirical evidence for the claim that individuals can have observational awareness that they act on implicit biases. A study from Monteith, Voils, and Ashburn-Nardo (2001) found facilitated response times to pairing tasks involving stereotypically congruent items (in this case, with black names and unpleasant terms, and white names and pleasant terms) than stereotypically incongruent items (white names with unpleasant terms, black names with pleasant terms). After the study, Monteith and colleagues questioned participants, and discovered that a significant proportion of them (64%) recognised that they were responding differently on the incongruent trials as compared with the congruent trials.

Participants were then asked to write down what might explain the difference in the timing of their responses. Their explanations were classified as appealing to factors which did implicate the participant in question as having biased racial or stereotypical associations, and those which did not.[25] Of the 64% of participants who were observationally aware of their discrepant responses, one third explicitly attributed them to racial associations that they personally held (but did not necessarily endorse), saying things like "One typically hears unpleasant words paired with Black names, especially on the news. Unfortunately, I felt automatically drawn to the pairing of the two" and "It was easier to pair pleasant words with your own race" (Monteith *et al.*, 2001: 408). The set of those who detected a discrepancy *and* attributed it to a racial association amounts to approximately 27% of the whole sample. The two thirds who did not mention racial factors in their explanations said things like "The Black names are unfamiliar and harder, so they went better with the unpleasant words" or because "Black is associated with the dark and scary negative things. White is associated with bright and happy things" (Monteith *et al.*, 2001: 408).[26] So, whilst it was by no means the majority who both detected *and* attributed their discrepant responses to a stereotypical racial association, this is persuasive evidence that it is at least

[25] Explanations were classified by two independent judges. There was agreement on 94% of the cases (Monteith et al., 2001: 408).
[26] One might think that these statements *do* imply stereotypical racial associations, but the point the experimenters seem to be making is that the participants themselves did not recognise them as racial.

possible to detect implicitly biased behavioural responses and to attribute these associations to oneself. Detection of the discrepant responses was also associated with feelings of guilt across the set of participants who were observationally aware of their discrepant responses, whether they attributed their responses to racial associations or not. One might think that feeling guilty is only appropriate when someone detects that they have done something wrong, even if they are unwilling or unable to articulate what the wrongness consists in.

Holroyd cites some more recent evidence that people can have observational awareness both that they harbour implicitly biased attitudes, and that their actions may be influenced by such attitudes. Hahn *et al.* (2014) asked participants to think carefully about how they would respond on an IAT test. As Holroyd points out, Hahn and colleagues discovered that individuals were able to predict their experimental performance: Participants' responses to questions such as "My sorting of [the congruent pairings] will be very/moderately/slightly easier…" corresponded with how they actually performed on the IAT. Participants also attributed the relevant associations to themselves: "My true implicit attitude is a lot/moderately/slightly more positive towards *white*," (Hahn *et al.*, 2014: 5-8; quoted in Holroyd, 2015: 519). Holroyd maintains that individuals were not merely reporting or using their explicit attitudes to make predictions here, because their predictions did not always match explicitly reported attitudes (Holroyd, 2015: 519-20). Further, she suggests that participants did not make their predictions on the basis of citing a general stereotype they took to be prevalent in society: there was divergence among individuals' predictions of the general societal stereotype, whilst predictions of their own attitudes correlated with their IAT performance, (Holroyd, 2015: 519). Holroyd maintains that:

> This study is important, because it indicates that individuals are not only able to detect morally relevant features of their actions post hoc; they were also able to predict morally undesirable features ex ante. …it is surprising in the context of philosophical discussions that have supposed that implicitly biased behaviour is something of which individuals are not (in some sense) aware, and have elided the different notions of awareness at issue. But these assumptions are not supported: there is evidence that supports the claims that, with reflection, individuals are at least sometimes able to detect and predict discrepant responses. (Holroyd, 2015: 520)

As Holroyd points out, these findings are surprising in the context of at least some philosophical discussions which have tended to assume that we lack any awareness of implicit bias. But she reiterates the point that we have also seen Payne and Gawronski (2010) stress—that the original distinction between the implicit and the explicit was never made on the basis of awareness. In light of this, the above results are perhaps less surprising but nonetheless equally significant for our purposes.

So, given that we may have observational awareness of at least some implicit biases (and their influence on our actions), there are no grounds for drawing a substantial distinction between implicit biases and beliefs, and their effect on our behaviour, on the basis of observational awareness.

3.2.2. *The SD thesis and inferential awareness*

It certainly seems possible for many individuals to gain inferential awareness of their implicit biases by acquainting themselves with the relevant empirical literature. However, SD theorists might argue that a major hurdle to gaining this sort of inferential awareness is that the majority of people do not have access to the relevant academic journals in which such research is published. Further, it might be suggested that one does not need to read any journals to know that they have beliefs, but arguably, one may need to read journals to become aware that they have implicit biases.

I think that continuum theorists can say a number of things in response to these remarks. Firstly, lacking access to academic journals certainly is a hurdle to finding out about empirical research on implicit biases. But it is no less of a hurdle to gaining inferential awareness of academic research on beliefs. Surely this also involves accessing, and understanding, material published in access-limited academic journals.[27] And so, whilst it might be the case that one does not have to read any journals to become aware that they have beliefs, this will not, by definition, be knowledge that they gained inferentially (but, either observationally or introspectively). Our question in this subsection is "is there any difference in

---

[27] For an example of a body of academic research on beliefs, see the 'Cultural Cognition' project based at Yale. This project has, for instance, assessed the US population's beliefs about climate change (for example, see Kahan, D. *et al.* (2011) 'Cultural cognition of scientific consensus', Journal of Risk Research, 14:2, 147-174) as well as having assessed the US population's political beliefs (for example, see Gastil, J. *et al.* (2011) 'The Cultural Orientation of Mass Political Opinion', *PS: Political Science & Politics,* 44:4, 711-714).

the *inferential awareness* that we have of our beliefs as compared to that which we have of our implicit biases?" So this possible response from the SD theorist is not appropriate to this section, and rather belongs in the discussion of observational awareness (where I have already refuted the SD view) or introspective awareness (where I will shortly refute the SD view). Further, given that there seems to be a recent rise in media coverage of implicit bias in the popular and business press, presented for non specialists, we might think that it is relatively easy to gain inferential awareness of academic research about implicit biases.[28]

SD theorists might respond by arguing that perhaps the definition of inferential knowledge as knowledge of a body of academic research is too specialised, and that we can have a bank of cultural knowledge from which we may infer that we have beliefs, but not that we have implicit biases. But I think that the notion of the unconscious, and the legacy of Freudian psychology is also sufficiently part of the bank of at least western cultural knowledge that relaxing the definition of inferential awareness in this way will not generate a substantial distinction between (western) cultural awareness that we have beliefs and cultural awareness that there are elements of our psychology much like implicit biases.

So, given that it is possible to gain inferential awareness of both some beliefs and some implicit biases, there are no grounds for drawing a substantial distinction between implicit biases and beliefs on the basis of inferential awareness.

3.2.3. *The SD thesis and introspective awareness*

For there to be a substantial distinction between agential attitudes and implicit biases on the basis of introspective awareness, it must be the case that (i) we lack introspective awareness of all of our implicit biases and their influence on our actions; and that (ii) we have introspective awareness of all of our agential attitudes, such as our beliefs, and their guidance of our actions. In the following, I will show that both (i) and (ii) are false.

---

[28] For examples of such coverage in recent news, see Frith's 'How can we stop unconscious bias', published on 28/11/15, available at: http://www.bbc.co.uk/news/science-environment-34910954 and Elsesser's 'Be Careful Hollywood: Raising Awareness Of Unconscious Bias May Backfire', published on 02/12/15, available at: http://www.forbes.com/sites/kimelsesser/2015/12/02/be-careful-hollywood-raising-awareness-of-unconscious-bias-may-backfire/

Claim (i) is false.

Borgoni (2015) raises a challenge against claim (i): that we lack introspective awareness of all of our implicit biases, and their influence on our actions. She argues that on her 'ordinary' notion of introspection, in which we direct our attention to the inner manifestations of our mental states, we *do* count as having introspective awareness of at least some of our implicit biases. Recall that on this account, introspection

> involves active self-reflection and analysis of the manifestations of one's mental states, such as thoughts, feelings, memories of one's actions in particular circumstances and other mental states related to the one in question. (Borgoni, 2015: 216)

Borgoni illustrates her claim with the example of 'Emilia'. Emilia is an academic in the humanities, who grew up in an environment where family and friends often asserted that women are not fit for politics. Despite this, Emilia has dedicated much of her research to gender issues, and, as a result, she now wholeheartedly rejects the assertions of family and friends in her childhood, and sincerely believes that the different genders are equally fit for political activity. Whilst Emilia's research has prompted her to adopt explicit egalitarian beliefs, and to sincerely avow egalitarianism in conversation with colleagues, it turns out that Emilia is "sexist in most of her unguarded, instinctive and automatic behavior when it comes to assessing female performance in politics" (Borgoni, 2015: 213). For instance, it is the case for Emilia that

> women's proposals and performance in political debates rarely seem as good as men's. When they do, she is amazed by this fact. Sometimes Emilia catches herself thinking that a given female politician will not succeed in her position, without having any evidence in support of her prediction. But then she quickly represses such a thought. Emilia is aware of the pattern underlying her automatic reactions and tries to correct their impact on her explicit attitudes and decisions. (Borgoni, 2015: 213)

Borgoni argues that Emilia is an agent with explicit egalitarian beliefs but with implicit attitudes with a somewhat contradictory content. For Borgoni, these attitudes count as implicit because they have not updated consistently with the

evidence that she has for egalitarian thinking, and they pop up in thought automatically, even though she has tried hard to suppress them in the past. Such features also render these attitudes as implicit for Gendler (2008a, 2008b, and as I summarised her in claims TG3 and TG4)[29] who stresses that resistance to evidence characterises implicit bias. Borgoni argues that on her ordinary account of introspection, Emilia counts as introspectively aware both of her having implicit biases against women in the political sphere, and of how these biases manifest in judgement:

> She knows how she feels when assessing the prospects of female candidates. She knows the sorts of thoughts she has when she listens to the discourse of female candidates. She has first-personal access to her phenomenology and is able to acknowledge that this is evidence of her continuing to hold a prejudicial belief. She is able to know that she still holds the prejudicial belief *from the inside.* Such access amounts to traditional first-personal access to her thoughts. (2015: 216)

In light of Borgoni's argument, two options are open to SD theorists who wish to maintain the distinction between agential attitudes and implicit biases on the basis of introspective awareness. They can either (A) reject Borgoni's 'ordinary' account of introspection, or (B) deny that Borgoni's description of the case is correct, and insist that the thoughts on which Emilia is introspecting are *explicit* agential attitudes after all. Option (A) involves committing to an account of introspection which builds in some extra criteria other than straightforward awareness, which agents like Emilia will not fulfil. As Borgoni (2015) suggests, one option here would be to adopt an account on which to count as introspecting, the agent must simultaneously judge the belief of which they are aware to be true (for example, see Moran's 2001 account). Even here, though, the continuum theorist might push back and argue that Emilia*,* in the instant that she finds herself thinking that a female politician will not succeed in her position, genuinely *does*

---

[29] These claims were summarised as follows:
TG3: Implicit biases, insofar as they are aliefs, are not sensitive to the propositional information encoded in mental states such as beliefs: Learning that not-*P* may well not cause me to cease having an implicit bias with the apparent content that *P*.
TG4: Beliefs change in response to changes in evidence, implicit attitudes change in response to changes in habit. If new evidence won't cause you to change your prejudiced behaviour in response to an apparent stimulus, then your reaction is due to an implicit attitude rather than belief.

judge it to be the case that the female politician will not succeed. It's just that she quickly goes back on this judgement immediately after she makes it.

Accepting option (B) might therefore be preferable for the SD theorist. However, this option is not open to all SD theorists. By at least some SD theorist's own lights, Emilia's automatic thoughts *cannot* count as explicit, agential attitudes because they fail to have other characteristics which those SD theorists think are necessary for agential attitudes. For instance, they are resistant to her evidence, which, as I just mentioned, would render these attitudes as implicit for Gendler (2008a, 2008b). They fail to respond to reasons that Emilia sees there to be for egalitarianism, and so fail to be integrated with her endorsed egalitarian beliefs. As such, they fail to meet the necessary criteria for agential attitudes for Levy (2014a).[30]

So, the thoughts on which Emilia introspects don't seem like characteristic explicit beliefs, because as soon as she recognises that she is entertaining these thoughts, she does not endorse them, and attempts to suppress them. They look more like mental states that might be found somewhere in the middle of the continuum, with unconscious, unendorsed, evidence-insensitive attitudes at one end, and conscious, endorsed, evidence sensitive attitudes at the other. So, Borgoni's (2015) argument presents something of a problem for SD theorists who claim that we have no introspective awareness of any of our implicit biases.

Claim (ii) is false

Claim (ii) is that we have introspective awareness of all of our agential attitudes and their guidance of our actions. Before presenting my main argument against (ii), it is worth pointing out that there is no consensus on (ii) across the philosophical community (Hume, 1748/1977; Ryle, 1949; Carruthers, 2011). For instance, for Hume, there are some agential attitudes which cannot be introspected, because being in an appropriate mental state to so introspect entails that we are no longer in those former attitudes which we intended to introspect in the first place, (Hume, 1748/1977: Introduction). Ryle summarises Hume's point in the following:

---

[30] In particular, these claims were summarised to:
NL3: Being integrated with each other is a necessary condition for mental states and their guidance of actions to be agential.
NL6: It is characteristic and perhaps even definitive of implicit biases that they do not respond to our reasons.

> There are some states of mind which cannot be coolly scrutinised, since the fact that we are in those states involves that we are not cool, or the fact that we are cool involves that we are not in those states. No one could introspectively scrutinise the state of panic or fury, since the dispassionateness exercised in scientific observation is, by the definition of 'panic' and 'fury', not the state of mind of the victim of those turbulences. (Ryle, 1949: 147)

It might be that these attitudes are available to *retro*spective awareness, but this opens up the potential for error, and mis-remembering (Ryle, 1949: page). So if one's chosen theory of introspection involves a faculty which is direct and immediate, then there are at least some personal level mental states which are not available to this faculty.

More generally, Ryle (1949) argues that theorising about introspective awareness is somewhat misdirected, and that really, we do no such thing as introspect when we want to discover the contents of our minds. Instead, according to Ryle, we observe our behaviour, or ask 'dispositional questions' about how would we respond in certain circumstances, and infer what we believe, or intend, or judge, and so on, from our observations (Ryle, 1949: 151). In a similar vein, Carruthers (2011) suggests that we employ a process of self-interpretation to become aware of our beliefs, intentions and judgements, rather than a process of introspection. These views, in effect, deny that there is any such thing as introspection, and maintain that we come to know about our personal-level mental states through the interpretation either of our actual actions, or those that we discover that we are disposed to perform in some hypothetical circumstances. If these views are right, and it is the case that our awareness of our own personal level mental states is observational (in Holroyd's terminology; inferential in Ryle's), and it is also true that we have observational awareness of at least some of our implicit biases, then there will be no distinction in kind between the awareness we have of our personal level mental states and of our implicit biases. The Ryle/Carruthers view is not the dominant one, however, and so I will not pursue it any further as a potential response to the SD theorist. But I think it is worth having this view in the background of the discussion to demonstrate that, if there is a distinctive faculty for introspection, then it is not immediately obvious as to what characterises it.

Holroyd (2015) presents another problem for claim (ii) which does not depend on the wholesale rejection of introspective awareness. She agrees with SD theorists Saul, and Kelly and Roedder that we don't have introspective awareness of our implicit biases or their influence on our actions, (2015: 516). However, she argues that individuals may act agentially even when they lack introspective awareness of the agential attitudes and commitments which guide the action in question. To illustrate this claim, Holroyd borrows an example from Snow (2006), in which a person 'instinctively' intervenes when she observes an elderly woman being cheated by a sales clerk (Holroyd, 2015: 516). Holroyd maintains that:

> A central feature of Snow's example is that the agent does not recognise that her sense of justice is activated (her justice related goals, in Snow's terms); there are important aspects of her cognitions, then, in relation to which she lacks introspective awareness. But that she lacks awareness of this aspect of her cognition does not mean that she cannot be held responsible and praised for her actions. (Holroyd, 2015: 516).[31]

Holroyd also talks about cases of forgetting, in which agents fail to have introspective awareness (or indeed any kind of awareness) of the cognitive processes which led them to forget, but in which, intuitively, the agents still seem to be morally responsible (and so, presumably, they also count as acting agentially). Holroyd appeals to Sher's *Hot Dog* case (originally in Sher, 2006: 286-7, see also Sher 2009) in which a person goes to pick up her children, and then has some unexpected issues to settle at school, forgetting that she has left her dog languishing in her increasingly hot car. Sher suggests, and Holroyd agrees, that even though the individual in question lacked introspective awareness of the processes that brought about her omission, she is not therefore exculpated of blame (and so we might also think that the omission should be thought of as agential). Speaking of the individual who defends the shop assistant, and the individual who forgets her dog, Holroyd argues that:

---

[31] I am assuming that Holroyd's argument in the above quotation depends on an implicit premise that an action may be agential, and correctly attributable to the agent (which presumably is a necessary condition on its being culpable) even though the agent in question does not have introspective awareness of the cognitive processes which bring about their action.

> …what these cases show is that *merely lacking introspective awareness of the processes* involved in deliberation and action does not suffice to exculpate, and is consistent with praiseworthy and blameworthy action. Nonetheless, the more familiar processes described above are similar in some important respects (whilst of course dissimilar in others) to those involving implicit associations that produce implicitly biased actions: they are fast, automatic, not readily under the agent's deliberative control, unreflective, and (in the latter case) processes the agent would not endorse, and productive of morally undesirable outcomes. Unless a case can be made for implicit associations being treated differently, then the fact that an agent lacks awareness of the operation of implicit associations would not be grounds for exemption from responsibility. (Holroyd, 2015: 516)

So, it appears that we cannot uphold a substantial distinction between implicit biases and beliefs, and the influence of each on action, with respect to introspective awareness: We may fail to have introspective awareness of the former, but we also sometimes fail to have introspective awareness of the latter.

However, this argument from Holroyd is open to an objection articulated by Levy in his account of consciousness$_{PA}$ (2014a): that *occurrent* introspective awareness is too stringent a criterion for a mental state to count as agentive: Holroyd's cases are those in which agents lack occurrent introspective awareness of the influence of their beliefs on their actions. That is, they are not aware that these attitudes guide their behaviour in the actual sequence of events. As we saw in the previous chapter, for Levy (2014a), being occurrently aware of the influence of an attitude on an action was too stringent a criterion for the attitude and the action which it guides to count as agentive. Accordingly, he argued that even if the agent is not occurrently aware that an attitude guides an action, as long as the guiding role that the attitude plays is effortlessly recallable, then the action may be agentive. Recall from Chapter 2 that, for Levy, an attitude is taken to be available for effortless recall if the presence of an ordinary cue for that attitude causes the agent to become occurrently aware of the attitude and the guiding role that it plays (Levy, 2014a: 34).

Holroyd maintains that it is not possible to have occurrent introspective awareness of an implicit bias and its influence on behaviour (2015: 516), and so it would seem that implicit biases will fail to be available for effortless recall, at least according to Holroyd's account. If Holroyd is right about this, then it would

seem to follow that no implicit biases have ever been occurrently tokened. If an implicit bias has never been occurrently tokened, then it cannot be recalled in the presence of an ordinary cue—indeed, Levy would likely argue that there are no ordinary cues for implicit biases and therefore that we are not conscious$_{PA}$ of our implicit biases.

If Levy is right, then we should conclude that there *is* a distinction between our awareness in the case of implicitly biased action, compared to at least some cases of agential action: Whilst both the shopper and the implicitly biased agent are not *occurrently* introspectively aware of the influence of their attitudes on their actions in the actual sequence of events, the shopper's justice-related attitudes may be *effortlessly recallable* for her (rendering her conscious$_{PA}$ of the attitude's action-guiding role) whilst an implicit bias is never effortlessly recallable (and so agents can never be conscious$_{PA}$ that implicit biases guide action).

However, I am not convinced that this new SD argument that relies on Levy's (2014a) account of consciousness$_{PA}$ works. In the remainder of this section I raise some problems for it. Specifically, I argue that (a) the notion of availability to effortless recall is vague, due to problems specifying what counts as an ordinary cue; and that even if we put that problem aside and characterise availability in terms of an unspecified but intuitive notion of what counts as an ordinary cue, (b) effortless recall of the guiding attitudes is not a necessary condition for an action to count as agential. In response to my objections, SD theorists may seek to make the distinction between the awareness that we have of beliefs and implicit biases, and the influence of each on our actions, by appeal to an alternative account to consciousness$_{PA}$. Anticipating this possibility, I will show how the problems that I raise for the distinction on the basis of consciousness$_{PA}$ will be inherited by alternative accounts.

Recall that on Levy's account of consciousness$_{PA}$, an attitude which brings about some action is considered to be a personal-level state, and part of their evaluative stance, if it fulfils two conditions: firstly, that of being online; and secondly, that of being effortlessly recallable (2014a: 34). An attitude is online if it is playing some role in guiding action. An attitude is effortlessly recallable if a large range of 'ordinary cues' would render the state occurrently conscious (in that the person in question would become occurrently aware of the attitude, or, in Levy's words 'occurrently token' its content, were they to encounter such cues).

According to Levy, no special prompting, such as asking a leading question, should be required for a state to be effortlessly recalled—if it is, then the state does not qualify as effortlessly recallable (2014a: 34). For instance, according to Levy, a telephone counts as an ordinary cue to prompt recall of the information that it is a friend's birthday. If someone does not recall that it is their friend's birthday in the presence of a telephone, then the information that it is their friend's birthday is not effortlessly recallable for them.

This account of effortless recall might seem intuitive. But on closer inspection, it becomes clear that the notion of what counts as an ordinary cue is underspecified. How are we to understand what constitutes an 'ordinary cue'? Is it the case that the same set of cues *C* count as 'ordinary' for all agents to recall some content *P?* For instance, is it the case that telephones *always* prompt recall of the fact that it is a friend's birthday for agents who are $\text{conscious}_{\text{PA}}$ that it is their friend's birthday? Or is there some agent-specificity as to what counts as ordinary from one agent to the next? It would seem that the former is a little too strong: surely telephones can only count as ordinary cues for recalling that it is a friend's birthday for agents who own and utilise telephones, which is by no means the case for *all* agents who remember birthdays. So it seems that the kind of items that count as ordinary cues for the recall of certain information are going to be at least partly specified by variations in different agents' environments. Given the diversity of cultural practices and substantial variation in the material environments of different agents the world over, there will be considerable inter-agent variation in terms of what counts as an ordinary cue for the easy and effortless recall of *P* from one agent to the next.

The SD theorist might respond to this concern in the following way: That there is inter-agent variation might not be too problematic, given that Levy's $\text{conscious}_{\text{PA}}$ is about *personal* availability of attitudes. It might mean that it is hard to tell, for any one agent, whether some informational state *P* is effortlessly recallable for *them*, until we know a considerable amount of information about their interaction with various objects in their environment. But even if it is hard for us to determine what counts as an ordinary cue for any one agent, that in itself doesn't mean that there isn't some fact of the matter as to which objects are, for *S*, ordinary cues for the recall of some informational state that *P*. It just means that there aren't easily knowable general facts about what counts as an ordinary cue for any one agent.

This might avert specificity concerns between agents. However, alongside inter-agent variation, I think there will also be considerable intra-agent variation. That is, there will be considerable variation in what counts as ordinary for the recall of some informational state *P* for one and the same agent at different times, depending on what else is currently occupying their attention. For instance, a telephone might prompt an agent to effortlessly recall that it is their friend's birthday in a circumstance in which they are relatively cognitively unoccupied, but if, in another circumstance, they are waiting for an important call from the hospital regarding the health of a relative, encountering the telephone might fail to prompt their recall of the relatively less important birthday information.

Indeed, it would seem that agents are frequently occurrently unaware that particular attitudes guide actions, even in the presence of objects that seem, intuitively, to be ordinary cues for recalling the mental states in question. A person getting ready in the morning may be deep in thought about the day ahead, and hence be occurrently unaware that beliefs about the benefits of tooth-brushing, as well as beliefs about the spatial location of their toothbrush and toothpaste guide their tooth-brushing actions. But if any object serves as an ordinary cue for recalling the mental states which guide tooth-brushing, then it seems that it would be either a toothbrush or a tube of toothpaste. Similarly, an audience member in a lecture might reach for a pen to write down an interesting point without being occurrently aware that their beliefs about the capacity of the pen to record information guide their reaching action, even though, if any object serves as an ordinary cue for recalling pen related beliefs, then it would seem to be a pen.

When it comes to the intervening shopper, we might ask, what sort of objects would have cued her to become occurrently aware of the justice related beliefs as they guided her action? Since she wasn't occurrently aware of those beliefs as they guided her action in the actual sequence of events, just encountering the person in distress that she acted to help wasn't sufficient to cue her recognition of the justice related beliefs as they guided the intervening action. But if the presence of the person in distress isn't an ordinary cue for the recall of the relevant beliefs in this instance, then it is hard to imagine what sort of object would serve this purpose in her case. So, now we have a putatively agential action that fails Levy's test for the necessary preconditions of awareness for agential actions, making his effortless recall condition seem too restrictive.

The SD theorist might try to respond by arguing for something like the following:

> If some ordinary cue *O* sometimes causes *S* to effortlessly recall (and so to become occurrently introspectively aware of) their belief that *P*, and that *P* guides action, but sometimes doesn't cause *S* to effortlessly recall their belief that *P*, and that *P* guides action, depending on what else is occupying their attention, then one and the same belief that *P,* and the role that *P* plays in guiding action, are *sometimes* conscious$_{PA}$ and *sometimes* not.

But the problem with this is that the actions of the intervening shopper, the tooth-brushing agent, and the lecture note-taking agent *do* look like cases of agential action, even though the agents in question are not caused to become occurrently aware of the attitudes which guide their actions in the presence of what seem intuitively to be ordinary cueing objects for these attitudes. So, even if we accept some intuitive notion of what counts as an ordinary cue, there is still considerable intra-agent variation as to whether the presence of an ordinary cue for a particular attitude actually does cause the agent to become occurrently aware of the mental state, and the guiding role that it is playing in a number of agential actions. So, that a mental state is available for easy and effortless recall does not seem to be a necessary condition for the action that it guides to be agential. So, as per (a): the notion of availability to effortless recall is vague, and the notion of an ordinary cue is underspecified, and as per (b): effortless recall of the guiding attitudes is not a necessary condition for an action to count as agential after all. As such, consciousness$_{PA}$ is too restrictive an account to characterise the kind of awareness that is necessary for agency.

Supporters of the SD thesis might think that these conditions are too stringent, and that perhaps a state's guidance of action only needs to be effortlessly recall*able* in the presence of an ordinary cue, but not necessarily effortlessly *recalled*. However, this would effectively be to abandon consciousness$_{PA}$ and return to an unrestricted notion of dispositional awareness, the very sort of account of awareness which Levy (2014a) suggested is too inclusive to ground agency and moral responsibility. For Levy, effortless recall in the presence of an ordinary cue is supposed to be an analysis of the already dispositional notion of consciousness$_{PA}$. It is supposed to restrict an attitude's disposition to become occurrently conscious. Accordingly, an attitude is

conscious$_{PA}$ only if it is available for effortless recall. A state is available for effortless recall if the presence of an ordinary cue causes the agent to become occurrently aware of the state and its role in action. That is to say: should an agent be in the presence of an ordinary cueing object, they *will,* in that sequence of events, effortlessly recall and so become occurrently aware of the mental state in question, and how it guides behaviour (Levy, 2014a: 34). To maintain instead that the presence of an ordinary cueing object renders the states which guide action effortlessly recall*able*, but not necessarily effortlessly recall*ed*, does not offer an analysis of what it is for a mental state to be conscious$_{PA}$. It simply introduces another disposition. If SD theorists do not specify the conditions under which agents *would* become occurrently introspectively aware of the mental states which guide agential actions, and rather maintain that there is just some unanalysable sense in which they c*ould* become aware of them, then we're back to an unrestricted dispositional account, the very account that the effortless recall condition of conscious$_{PA}$ was designed to improve upon.

At this point, an SD theorist might accept that Levy's requirement that the states which guide actions are available for effortless recall is not a necessary condition for the actions in question to be agential. But they might maintain that there is still a difference in what agents *could* be introspectively aware of in the case of tooth-brushing and lecture note taking actions, and implicitly biased actions. Here, SD theorists may argue that agents who prepare to brush their teeth and who reach for pens, without occurrent introspective awareness of the attitudes which guide these actions, have, at least at *some* point, been occurrently introspectively aware of the mental attitudes which guide action. In contrast, agents have never been occurrently introspectively aware of their implicit biases, or how they guide actions. Because of this, it may be argued, implicit biases are not even in principle recallable. The relevant difference in awareness, then, is that agential actions are guided by states of which the agent has at some point been aware, even if they are not occurrently introspectively aware of the role that these states play in the present action, whilst implicitly biased agents have never been occurrently introspectively aware of the attitudes which guide their implicitly biased actions.

However, I do not think that this response enables SD theorists to distinguish between *all* agential attitudes and their guidance of agential actions, and implicit biases and their guidance of implicitly biased actions. This is because

there seems to be many cases in which agents acquire attitudes, which then go on to guide what look very much like agential actions, without the agents in question being occurrently consciously aware of the guiding role of these attitudes. Consider the following cases.

> *Lights:* Muhammed is tidying the kitchen one evening when, all of a sudden, the lights go out. Before he has time to reflect on what he is doing, he has reached up through the darkness to find the handle to the cupboard where the torch is kept, even though he was not occurrently introspectively aware of forming a belief about where the cupboard door was in relation to him before the lights went out.
>
> *Flat warming:* Laura goes to Aisha's flat warming party. When she gets to Aisha's building, she glances at the sign to number 22, and ascends two flights of stairs, whilst writing a long text message to another friend about how to get to Aisha's building, before arriving outside flat 22, Aisha's door. Later on, Aisha bemoans the lack of a lift in her building, to which Laura replies "You're lazy if you want a lift to the second floor!" Laura says this even though she was not occurrently consciously aware of forming the belief that Aisha lives on the second floor, nor of how this belief guided her action as she climbed the stairs to Aisha's flat.
>
> *Mushrooms:* Naveen returns from the shops and unpacks the shopping. Julia asks "did you buy any mushrooms?" and Naveen replies "yes, bottom shelf of the fridge, in the black punnet." He performs this utterance automatically, even though he wasn't occurrently introspectively aware of the colour of the container as he unpacked the shopping—instead he was in deep thought about those revisions to his latest paper. Nevertheless, he formed a belief about the colour of the punnet without occurrent awareness, a belief which went on to guide his utterance without his occurrent awareness.

In these three examples, agents are not occurrently introspectively aware of either forming an attitude, or of having an attitude, before it goes on to guide action, which it does without the agent's occurrent introspective awareness. However, Muhammed's reaching action, Laura's successful navigation to Aisha's flat, and Naveen's utterance to Julia nonetheless all look like examples of agential actions. So, the requirement that an agent has to have been occurrently introspectively

aware of an attitude at some point before it goes on to guide an action will not distinguish *all* agential actions from implicitly biased actions. Instead it distinguishes some agential actions from some other agential actions and implicitly biased actions. That is not enough to establish the required substantial distinction.

To sum up my argument against SD claim (ii) that we have introspective awareness of all of our agential attitudes, and their guidance of our actions, I have argued that: (1) It is not necessary that an agent should be occurrently introspectively aware of an attitude and how it guides action, for the action to be agential; (2) It is not necessary that an agent should effortlessly recall, and thus occurrently introspect, the action guiding role of an attitude in the presence of an ordinary cue for that attitude, for the action to be agential; and (3) it is not necessary that an agent should ever have been occurrently introspectively aware of an attitude, or its guidance of behaviour, for an action guided by that attitude to be agential.

So, contra SD claim (i), I conclude that we sometimes *do* have introspective awareness of our implicit biases and their guidance of our actions (if one is convinced by Borgoni's 2015 account); and contra SD claim (ii), I conclude that we sometimes *lack* introspective awareness of our agential attitudes, and their guidance of our actions. So, the SD claim as regards introspective awareness is false, and does not help to draw the required distinction.

3.3. Implicit Biases and Observable Class Preferences

Let us take stock of the argument so far. We have seen that SD theorists may not generate a substantial distinction between implicit biases and beliefs (and the influence of each on action) on the basis of any of the senses of awareness relevant to the implicit biases literature: observational, inferential, and introspective awareness. We have the same kind of observational, inferential and (at least on Borgoni's account) introspective awareness of at least some of our implicit biases, and their influence on action, as that which we have of at least some of our beliefs and their influence on action. For those unconvinced by Borgoni's account, there is still no substantial distinction between implicit biases and beliefs (and the influence of each on action) on the basis of introspective awareness, because beliefs may guide actions agentially, even when the agent has never been occurrently introspectively aware of the belief in question.

In this final section, I offer a positive account of awareness of implicit bias. I do so by introducing the notion of what I call an 'observable class preference' and arguing that both everyday agential attitudes and at least some implicit biases constitute observable class preferences. I argue that, with reflection, we are able to become aware of at least some of our implicit biases insofar as we may reflect on our preferences and observe them. If so, and relying on Borgoni's (2015) account of introspection, it follows that we will count as having introspective awareness of our observable class preferences, including those which are implicit biases, in this way.

Of course, it is open to my opponents to reject Borgoni's account of introspection. In that case, the conclusion would be restricted: we will only have observational awareness of our implicit biases which are observable class preferences. However, this is not sufficient to generate the kind of substantial distinction the SD theorist requires, because crucially, it will follow that we *also* only have observational awareness of our everyday, agential class preferences too. Therefore, whether or not one accepts Borgoni's account of introspection, my conclusion stands: we have the same kind of awareness of those implicit biases which are observable class preferences as we do of at least some of our everyday, agential observable class preferences.

3.3.1. *Introducing observable class preferences*

Consider the following case:

> *Trees*
>
> Sara walks past a willow tree on the way to work. She thinks to herself that the willow tree has a graceful shape. She finds it quite aesthetically pleasing. Accordingly, it seems correct to say that Sara believes that 'the willow tree has a graceful shape'. Sara also likes to spend time in the local park with her grandchildren. She often takes her camera, and particularly likes the photos where the children are playing beside a row of mature cedars. With their striking spread of branches, Sara likes the way that the cedars look. It seems correct to say that Sara believes that 'the cedars are striking'. When Sara was looking for a new place to live, she chose a flat with a view of some silver birch trees. She liked the idea of being able to see their elegant silhouettes against the morning sun whilst she drinks her coffee, finding their outlines

pleasing. It seems correct to say that Sara believes that 'the silver birches are elegant'.

I am going to argue that the agent in the example 'Trees' has what I will call an 'observable class preference'. To have an observable class preference, an agent must meet the following two conditions:

Observable Class Preferences

(1) Introspective awareness of some singular manifestations, where:
(1a) *S* evaluates *a* as *p*
(1b) *S* evaluates *b* as *q*
(1c) *S* evaluates *c* as *r*
...etc

(2) A general preference is inferable from (1):
(2a) *a, b* and *c* are all instances of some class of *F*s
(2b) *p, q* and *r* are all qualities of the same class and valence

In order to fulfil (1), *S* will have undertaken some evaluative judgement that *a* is *p*, (and that *b* is *q*, and that *c* is *r*), and, accordingly, formed a belief with the content that *a* is *p*, (and that *b* is *q*, and that *c* is *r*) of which she is introspectively aware. The items *a, b* and *c* may be a broad range of entities. They might be objects, but they might also be more abstract items such as pieces of music, or films. They might be events, such as holidays, or festivals, or they might be activities, such as cycling or playing football. *a, b* and *c* do not necessarily need to pick out entities as a whole, they might instead pick out particular characteristics of an entity.

As regards (2), I rely on an intuitive notion of what it takes for a set of entities to constitute a 'class', but I take it that all of the examples of classes that I will use in this section are relatively uncontroversial. For example, Hanukkah, Eid al-Fitr and Christmas are all instances of the class 'religious festivals', whilst football, hockey and rugby are all instances of the class 'team sports'. I will be talking about grouping people into classes as regards social identity characteristics which are apparent to an observer. In doing so, I am not committed to the notion

that this exercise picks out properties which are essential to the identity of the people in question, as opposed to picking out socially constructed characteristics available to an observer. More broadly, in assuming that the notion of a 'class' is intelligible, I am not committed to any particular metaphysics—realism, for instance—about classes or the properties that serve to identify their members. Classes as I am using them here may also be social constructions, or useful fictions, and so on.

Regarding (2b), I take qualities to be characteristics that the agent judges an entity to have. It might be that these qualities are properties which the entity in question actually instantiates, or it might be that they are apparent qualities of the entity that are experienced by some agents, but not others. Again, I rely on an intuitive notion on what constitutes a class of qualities. For instance, 'tasty', 'delicious' and 'flavoursome' would all count as examples of the class of gustatory qualities. This particular set of qualities also shares the same (positive) valence, which is necessary for the fulfilment of (2b).

In outlining the notion of an observable class preference, I specify that evaluations are made about three distinct entities, although it might well be that the agent has made more than that number of evaluations. I think that in some cases, it may be sufficient to have just two evaluations for a class preference to be observable. However, for the following examples, I will consider that at least three evaluations are made. Finally, for the kind of phenomenon that I have in mind, it is not necessary for the agent in question to have actually realised (2), for a class preference to be nonetheless *in principle* observable for that agent.

With this in place, I argue that, in the 'Trees' example Sara counts as having an observable class preference. Here is how Sara meets the conditions.

(1) Sara has introspective awareness of the following evaluations:
The willow tree has a graceful shape
The cedars are striking
The silver birches are elegant

(2) A general preference is inferable from (1), because:
Willow, cedars and silver birches are all kinds of tree
'Graceful', 'striking' and 'elegant' are all positive aesthetic evaluations

From this, it is inferable that Sara has a positive aesthetic preference for trees, as a class of objects. It might be that there is a particular feature of trees, their organic fractal pattern, for instance, that Sara is drawn to. Trees, as the objects of her evaluations, are a sufficiently similar set of entities to be considered members of a class. Meanwhile, 'graceful', 'striking' and 'elegant' are all positive aesthetic evaluations. So, it is inferable that Sara has a general aesthetic preference for trees.

I am going to suppose that Sara has not yet become aware of her observable class preference. That is, she has not yet inferred from the fact that she believes that THE WILLOW TREE HAS A GRACEFUL SHAPE, THE CEDARS ARE STRIKING, and THE SILVER BIRCHES ARE ELEGANT that she has a general positive aesthetic preference for trees. This seems to me to be plausible. That is, it seems plausible for Sara to have made each evaluation on separate occasions with enough distance between each episode of belief formation for her to not have recognised the similarity between the beliefs she ends up forming. There isn't anything conceptually problematic about Sara, as described. In short, Sara has an as yet unobserved, but nonetheless, in principle, observable class preference.

That said, it seems quite possible for Sara to *become* aware of her class preference for trees. Perhaps looking at the photo of her grandchildren playing beneath the cedars prompts her to reflect on other trees that she has made aesthetic judgements about. Perhaps she thinks about her walk to work, and about the willow that she passes, which leads her to reflect on her experiences of other organic forms, and to realise that they too have been positive. Whichever it is, it seems that it is not beyond Sara's capacities to reflect on one of her evaluations, to recognise both the entity and kind of evaluation that she has made, and to then reflect on other instances where she has encountered that entity and made an evaluation of the same kind. That is, it is possible for Sara to realise that she has a general positive aesthetic preference for trees. Even if she never had that moment of reflection in the actual sequence of events, it is still the case that she has an in principle observable class preference.

Does Sara discover her class preference through introspective or observational awareness? Sara is not directly acquainted with the general class preference, at least not when she first realises it, because making the discovery involved observing commonalities between her existing attitudes, and using them

as evidence to infer that she has this more general attitude towards trees as a class. So it might be argued that this is not an exercise of introspective awareness, but of observational awareness. Recall Holroyd's (2015) examples of agents who observe their stereotypical behaviour on experimental trials, and who, from this, infer that their stereotypical attitudes must be guiding this behaviour, (from Monteith *et al.*, 2001; Hahn *et al.,* 2014).

But it seems that Sara is doing something different to the agents in the Monteith *et al.* (2001) and Hahn *et al.* (2014) experiments. She is not observing her behaviour and inferring, on the basis of what she observes, that she must have certain attitudes. In fact, she is doing something that a third person would be unable to do: She is using her own attitudes as evidence to make first-personal ascriptions. According to Borgoni's 'ordinary' notion of introspective awareness, this counts as introspection. The introspective process involves not only calling to mind certain attitudes, but using the attitudes that one calls to mind as evidence to learn about the attitudes that one has. As Borgoni puts it:

> …relying on evidence of oneself holding a certain belief to know about one's own psychology does not need to lead to third-personal belief attributions. Introspection, a standard first-personal method of knowing one's own beliefs, seems to involve reasoning from the evidence of oneself holding certain beliefs. (2015: 215)

So, at least according to Borgoni's ordinary notion of introspection, Sara counts as introspectively aware of her general aesthetic preference for trees, even if the outcome of her introspective reflection is a discovery for Sara herself.

I think that cases like Sara's are relatively common. Consider the following examples:

> *'TV'*
> Reflecting on her TV watching preferences, Naomi realises that she's started watching more nature documentaries. Nature documentaries are something that Naomi remembers, at least when she was much younger, being very bored by. However, in the last couple of years she has watched all of *Frozen Planet*, *Life* and *Blue Planet*. As she reflects on this, she realises that she now finds nature documentaries captivating in a way that she never used to.

'Vegetarian'

Since becoming a vegetarian, Tom has learned that south Asian food (and the anglicised versions of that cuisine) contain a lot of meat-free options. Tom has always liked the tangy, lentil-based dhansak. Over time, two other dishes that emerged as favourites for him are the piquant ceylon and the hot, sharp pathia. Tom reflects on the recipes for these dishes and realises that they all contain souring agents (tamarind, vinegar and citrus juice accordingly) which give them the hot and sour flavour that he likes.

'Festival'

Clare loves drum and bass music. She discovered this after she went to her first music festival, where she looked forward to discovering some new artists. When she got home, she reflected on the artists that she particularly enjoyed, so that she could buy their records. These artists included Joanna Syze, London Elektricity and DJ Storm, who, as it turns out, are all drum and bass producers. Clare recognises the fast pace, the dropped third beat, and the bass driven sound which characterises the music of all three artists, and embraces her newly discovered preference for this genre.

The agents in the three examples above have, on separate occasions, made similar evaluations about different entities which turn out to belong to the same class. Like Sara, they reflect on the individual evaluations, and realise that they feature a particular class of entities which are evaluated as having a particular set of qualities. That is, all of these agents reflect on their mental states to discover a general class preference for a set of entities. All of these agents engaged in some self-reflection to discover their class preferences. However, even if they did not take the time to do this in the actual sequence of events, this does not mean that their class preference is not, in principle, observable for them. Here are all the individual evaluations, and the class preferences which they constitute, from the above examples:

	Individual evaluations	**Observable class preference**
Trees	That willow is graceful The cedars are striking The silver birch is elegant	Trees are aesthetically pleasing
TV	I would watch *Blue Planet* again I enjoyed *Life* I thought *Frozen Planet* was beautiful	Nature documentaries are enjoyable
Vegetarian	I have always loved a dhansak I enjoy a pathia I like ceylon	I enjoy hot and sour curries
Festival	Joanna Syze has a great sound I enjoyed dancing to London Elektricity DJ Storm was great	I like drum and bass music

The above examples show that observable class preferences are a common feature of our everyday experiences of evaluating the merits of various entities. Agents evaluate the merits of one entity, and then go on to evaluate another entity from the same class in much the same way, without necessarily remembering how they evaluated the first entity, and without necessarily realising the commonalities between the evaluations. They may then introspect on the individual evaluations, and infer from these evaluations a more general class preference which they have. This attribution of a more general class preference to the self may count as an act of introspective awareness if one adopts Borgoni's (2015) account. Or, it may count as an act of observational awareness if one holds an account which rules out inferences made on the basis of the contents of mental states as introspective. (The utility of this point will become clear in the next section, when I compare everyday observable class preferences to implicit biases and show that there is no difference in the kind of awareness that we have of each.)

Regardless of whether the acknowledgement of a class preference is introspective or observational, once an agent has acknowledged that they have a class preference, it may function like a typical agential attitude. The agent might feel assent to the general class preference, and they might be disposed to assert the propositional contents of the general attitude in appropriate circumstances. For

instance, Sara might say "I find organic forms quite beautiful". It also seems possible that, on recognising a general class preference, an agent might resent or even reject it: For instance, Clare might resent the realisation that she finds drum and bass catchy, because her friends (who are all into electro swing) think that drum and bass is awful and should have died in the late nineties. Recognition of a class preference might also guide behaviour. For instance, Naomi's recognition that she now likes nature documentaries might guide her to purposefully search for new programmes of that genre when watching TV.

3.3.2. *Implicit biases manifest as observable class preferences*

In the following, I demonstrate that at least some implicit biases manifest as observable class preferences. Whether or not implicit biases have content, and whether that content is propositional, is the subject of the next chapter (where I will argue that at least some implicit biases *do* have propositional content, and that this gives us reason to doubt that other implicit biases are not propositional). For the time being, however, and regardless of what one thinks about the content debate, I think that we can still make some comparisons between implicit biases and less controversially propositional attitudes like beliefs, with regard to the way in which they influence our evaluations. I argue that we have the same kind of awareness of at least some of those implicit biases which are observable class preferences as that which we have of some of our everyday observable class preferences, such as those outlined in the examples of 'Trees', 'TV', 'Vegetarian', and 'Festival'.

As an example, imagine the case of Ben. Ben works for the police, and is often involved in promotion decisions. Ben takes himself to endorse equal opportunities, and thinks that people ought to be hired on the basis of merit, and certainly not on the basis of their gender. As such, Ben ticks the boxes to qualify as an agent who does not have explicit prejudices when it comes to hiring women. However, Ben does harbour implicit biases against women, which regularly influence his hiring decisions. This seems like a plausible scenario. Recall from Chapter 1 Uhlmann and Cohen's (2005) experiment in which people prefer to hire a male candidate over a female candidate for police chief, even though the candidate profiles show that they are equivalently skilled: Participants presented with a streetwise candidate, who has a male name, and an educated candidate who has a female name, prefer the 'streetwise' candidate, whilst participants presented

with the same streetwise candidate, but this time with a female name, and the same educated candidate, but this time with a male name, tend to prefer the 'educated' candidate.

Suppose that Ben receives an application for promotion from Louise, an officer who has clocked a lot of hours on street patrols, made a lot of arrests, and been involved in a number of high-risk raids. Whilst Ben is impressed by Louise's credentials, the fact that Louise has spent a lot of time on the street also indicates to him that Louise has not been involved in the more cerebral aspects of criminal investigation which happen in the office: piecing together evidence, liaising with forensic investigators, developing potential avenues for further enquiry, and so on. Because of this, Ben forms the belief that LOUISE IS NOT THAT SMART, and decides against promoting her.

This also seems plausible. Recall that Uhlman and Cohen's (2005) participants justified their implicitly biased hiring decisions by citing what they saw to be the strength of the preferred candidates or the weaknesses of the non-preferred candidates. Indeed, Sandis (2015) suggests that agents whose hiring decisions are influenced by implicit biases nevertheless do have agential reasons, reasons of which they are occurrently aware at the time of the decision. According to Sandis (2015), for a person with an implicit gender bias, the fact that an applicant is female renders the features which count against her as *more* salient than such features otherwise would be if the candidate in question was a man. So, it is not the case that an implicitly sexist hirer is aware of nothing whatsoever when they reject an application from a woman, and that they can tell no story about why they did so. What they are aware of is that some negative features count against hiring a female applicant. What they are unaware of is that these negative features are *more* salient to them when the applicant is female, compared to when the applicant is male. So, it's not the case that Ben isn't aware of *anything* when he rejects Louise's application for the promotion. What Ben believes, and what he is occurrently introspectively aware of in this scenario, is that LOUISE IS NOT THAT SMART.

Ben is regularly involved in promotion decisions. He receives a number of applications from candidates who have a lot of patrol experience, but less investigatory experience. Some street-wise applicants are promoted, others are rejected, even though they have similar backgrounds and qualifications. Over time, a clear pattern emerges: Ben significantly favours applications for

promotion from men over those from equivalently qualified women. Ben rejects Priya's application, even though she is just as qualified as many of the (male) applicants whom he has promoted, and does the same to Abby. At the time of rejecting these applications, he is occurrently introspectively aware that PRIYA IS A BIT DENSE and ABBY IS NOT VERY CLEVER.

Ben has now made three separate evaluations, and now believes the following three propositions: LOUISE IS NOT THAT SMART, PRIYA IS A BIT DENSE and ABBY IS NOT VERY CLEVER, (although he has not yet been occurrently aware of these beliefs all at the same time). Louise, Priya and Abby are all women police, and Ben has evaluated all of them (negatively) on the basis of their intelligence. Accordingly, Ben meets the criteria as outlined above to have an, in principle, observable class preference, as follows:

	Individual evaluations	*Observable class preference*
Ben	Louise is not that smart Priya is a bit dense Abby is not very clever	Women are not that smart

It seems possible for Ben to observe his class preference. Suppose that Ben was having a particularly self-reflective day, and was thinking about the people that he had promoted over the past year. Ben might think about Louise, and how she (appeared to him to be) not very smart. This might get him thinking about other women who have applied for promotion. He'd possibly then remember Priya and Abby, and his similarly negative evaluations of them. Ben could then call to mind all of the beliefs he formed about Louise, Priya and Abby at once, being then occurrently introspectively aware that LOUISE IS NOT THAT SMART, PRIYA IS A BIT DENSE and ABBY IS NOT VERY CLEVER. From this, he may infer that women, or at least, those in the police who have applied for a promotion, are just not that smart. That is, he may take these three attitudes as evidence for forming a new attitude, that WOMEN ARE NOT THAT SMART, just as Sara, Naomi, Tom and Clare did in the examples of everyday observable class preferences.

As regards whether Ben's attribution of a new attitude to himself is an act of introspective or observational awareness, the same point stands as it did for the everyday cases. If one's chosen theory of introspective awareness permits the attribution of an attitude to the self on the basis of what one infers from one's

existing attitudes, then Ben is introspectively aware of his class preference. If one's chosen theory of introspective awareness does not permit this, then Ben is observationally aware of his class preference. Either way, Ben has the *same* kind of awareness of his implicit bias as a general class preference as Sara, Naomi, Tom and Clare have of their everyday observable class preferences. Even if Ben never in fact has such a moment of reflection, in which he occurrently tokens the relevant proposition, that WOMEN ARE NOT THAT SMART, it nevertheless seems *possible* that he could. Accordingly, it is as possible for him as it is for Sara, Naomi, Tom and Clare to become occurrently aware of their everyday class preferences.

I think this makes a strong case for my claim that at least some implicit biases are observable class preferences, and that we have the *same* kind of awareness of them as we have of our everyday observable class preferences. But I recognise that one may object to my account of implicit biases as observable class preferences in a variety of ways, namely, on the basis of (1) agential endorsement, (2) bias recognition and (3) extent of awareness. In the remainder of this chapter, I answer three such possible objections.

3.3.3. *Answering objections to the observable class preferences account*

Objection 1: *Endorsement*

In the everyday cases, the agents in question *endorse* the class preference that they infer on the basis of the individual evaluations that they have made. By 'endorse' I mean that they take the proposition that the preference implies to be the true. For instance, Sara takes it to be true that TREES ARE AESTHETICALLY PLEASING, whilst Naomi takes it to be true that NATURE DOCUMENTARIES ARE ENJOYABLE, and so on. And even if Clare *resents* the fact that she likes drum and bass music, because all of her friends think it's terrible, there is still a sense in which she endorses I LIKE DRUM AND BASS MUSIC insofar as she takes it to say something true. Ben, however, takes himself to endorse equal opportunities, and to not favour people for promotion on the basis of their gender. As such, it might be argued that Ben would *reject* the proposition implied by his observable class preference, taking WOMEN ARE NOT THAT SMART to be false. If that is the case, then it might be argued that this rejection would prevent him from attributing the proposition to himself. So, there is a difference between everyday observable class preferences and those implicit biases which are observable class preferences, in that agents

endorse the former, but not the latter, and this may then prevent them from recognising the attitude as their own—or so my opponent might argue.

Firstly, it is not clear that rejecting the proposition implied by an attitude prevents one from attributing that attitude to oneself. This was evidenced in the Monteith *et al.* (2001) studies, where at least some participants did consider their actions to be explainable by their harbouring racial associations on some level. Further, Borgoni (2015) argues that Emilia (the agent who catches herself thinking biased thoughts about women in politics) is able to attribute her biased attitudes to herself, even though she is unable to endorse their contents. So, just because an agent does not endorse the contents of an attitude, it does not follow that they are prevented from being introspectively aware of that attitude as their own.

Secondly, I think it is possible that Ben *would* endorse the proposition implied by his observable class preference. After all, Ben thinks that he's got good reasons for thinking that women are not that smart: when he evaluated Louise's application, he concluded that she is not that smart. The same is true for both Priya and Abby. In fact, the reasons that Ben takes himself to have for thinking that women are not that smart are the same reasons that Ben took himself to have for making his evaluations of each of the women in the first place, along with the recognition that the conjunction of these evaluations implies a general preference. So, given that Ben already thinks that Louise, Priya and Abby are not that smart, it seems at least possible that he will endorse the notion that women are not that smart, on observing his class preference.

SD theorists might find this response unsatisfactory, because Ben has not recognised what seems like a crucial fact in this instance, namely that he is *biased* in believing as he does. I address this concern in the following.

Objection 2: *Recognising bias*

My opponents may argue as follows. Whilst Ben might introspect on the beliefs he formed when evaluating Louise, Priya and Abby, and use these as evidence to endorse WOMEN ARE NOT THAT SMART, he still remains unaware of a crucial fact, namely that he is *not justified* in evaluating Louise, Priya and Abby as he did. That is, he might recognise the proposition that is implied by his evaluations of the women, but fail to recognise the role of his implicit bias in shaping these evaluations. So, the inability to recognise the fact that his evaluations are

influenced by an implicit bias is the crucial distinction between the awareness that Ben has of his observable class preference and that which we have of our everyday observable class preferences. One might then argue that insofar as Ben does not recognise that his class preference is *biased*, it is not agential.[32]

First of all, I do not think that it is impossible for Ben to recognise that his general hiring preferences are biased (though this may be incompatible with his endorsing the proposition implied by the general preference, but as I said above, I do not think endorsement is necessary for attributing an attitude to oneself). He could certainly discover his bias by comparing the C.V.s and qualifications of his successful applicants with those of his unsuccessful applicants, observing that he has hired many more men with exactly the same qualifications as many of the women that he rejected. Here, he is using the C.V.s as prompts: none of the information that he uses to discover his bias in this case is, strictly speaking, new to him—all of it has been occurrently tokened at some point, when he originally assessed each applicant, even if he forgot much of this information shortly afterwards. So, it might be that in comparing the C.V.s, he discovers his bias in an act of observational awareness, but this is only because his memory is limited. If he could remember all of the beliefs he formed about each applicant, then he could introspectively discover his bias, at least in Borgoni's (2015) sense, without relying on the C.V.s, as he would be able to introspect on his attitudes about each applicant's qualifications, and compare these to their gender. So, it is possible for Ben to discover that he is biased in his general class preference.

Secondly, I don't see why it is necessary for an agent to *recognise* that a particular attitude is biased in order for that attitude to be agential. To see this, consider cases of *explicit* prejudice in which the agent in question both assents to, and asserts, a general class proposition which couples a social group and a stereotypical trait. For example, consider Katie, who believes that Muslims are violent, a belief which regularly guides her speech acts and interactions with people that she takes to be Muslim. Katie is systematically biased, in that she has formed a belief about a class of people on the basis of a very small group of individuals, a group of people that many Muslims argue are not representative of Muslims more generally at all. In fact, being systematically biased in this way is

---

[32] This is the sort of argument we saw Levy (2014a) make in Chapter 2. In Levy's terminology, Ben is not conscious of the facts that make his attitude morally significant, and so this attitude is not integrated with his evaluative stance, and should not be considered an agential attitude.

often a central characteristic of prejudice—that's just what prejudice is, a belief which isn't borne out by the evidence. But crucially, it does not seem necessary for Katie to acknowledge herself as biased in order for her to qualify as sufficiently introspectively aware of her prejudiced attitude for that attitude to be considered agential—as an attitude that it is correct to attribute to Katie, the agent. So, if explicitly prejudiced people do not have to recognise that they are systematically biased in order to count as sufficiently introspectively aware of their explicit prejudices for such prejudices to be agential, then even if Ben does not recognise that he is systematically biased, this does not undermine his having sufficient awareness of his general class preference for it to be considered agential.

Objection 3: *Discerning Extent*

One might argue that even if Ben observed his class preference, and became aware both that he thinks that women are not that smart, and that he is biased in doing so, still he would not know the *extent* to which this preference influences his evaluation of any one woman. My opponent might maintain that this is not the case for everyday observable class preferences, where we are introspectively aware of the extent to which the preferential class attitude affects our evaluations. So, it might be argued, Ben may be able to observe his class preference, but, unlike agents with everyday class preferences, he cannot discern the extent to which it influences his decisions and actions, and this undermines the extent to which the attitude, and the decisions that it influences, are agential.

However, it is not the case that all agents with everyday observable class preferences are able to discern the extent to which their preference affects their evaluations either. Imagine that Sara, the agent who has a general aesthetic preference for trees, visits an art installation by an artist who paints a lot of trees. Sara finds one of this artist's paintings particularly aesthetically pleasing. She recognises that she generally finds the fractal shape of the trees aesthetically pleasing, but she also likes the distinctively artistic features of the painting, such as the colours, the materials used and the composition. In this case, it is not clear that Sara is able to discern the extent to which her positive aesthetic judgement is influenced by her general preference for trees, *vs.* the extent to which it is influenced by the specific artistic properties of the painting. So, because at least some agents with everyday observable class preferences are not able to discern the

extent to which they influence decisions and actions, that Ben cannot do this is no basis for a substantial distinction.

SUMMARY

In the forgoing, I argued that there is no significant distinction between (i) implicit biases, and the actions that they influence; and (ii) agential attitudes such as beliefs, and the actions that they guide, on the basis of the kind of awareness that we have of each. I argued that we have the same kind of awareness of at least some of our implicit biases and their influence on our actions, as that which we have of at least some of our beliefs, and their guidance of our actions.

Specifically, I investigated whether any substantial distinction may be upheld on the basis of each of the three notions of awareness (inferential, observational and introspective) that Holroyd (2015) demonstrates are at issue in the literature on implicit bias. I argued that, at least sometimes, we can have the same kind of inferential, observational and introspective awareness of at least some of our implicit biases, and the actions that they influence, as that which we have of at least some of our agential attitudes, and the actions that they guide. I then gave a positive account of the awareness that we have of our implicit biases, arguing that at least some implicit biases are observable class preferences. I argued that, insofar as awareness qualifies an attitude as agential, these implicit biases, and their manifestation in action, ought to be considered at least as agential as many of our everyday observable class preferences, and their manifestation in action.

This result undercuts SDR arguments (SD arguments regarding moral responsibility) which rely on the claim that we do not have the kind of awareness of our implicit biases, and their influence on action, that renders such attitudes and actions as agential: I showed arguments for this claim fail. In light of this, consider the following claim, which formed an implicit premise of Saul's (2013) argument:

> **JS4:** It is a necessary condition for moral responsibility for having a mental state *m*/for action influenced by a mental state *m* that the agent is introspectively aware of *m*/that *m* influences action.

To the extent that I argued that we are introspectively aware of at least some of our implicit biases and their influence on action, S4 is *consistent* with our having moral responsibility for at least some implicit biases/at least some actions influenced by implicit bias. That it is consistent of course does not yet show that we *are* morally responsible for harbouring any implicit biases, or for any implicitly biased actions—just that this isn't ruled out.

Now consider Levy's (2014a) SDR claim:

> **NL1:** …only when we are conscious of the facts that give our actions their moral significance are those actions expressive of our identities as practical agents and do we possess the kind of control that is plausibly required for moral responsibility, (2014a: 1).

In §3.2.3, I argued that it is not necessary for our attitudes to be conscious_{PA} (according to Levy's (2014a) particular account of consciousness as personal availability) in order for them to qualify as agential, and in order for them to guide agential action. For Levy, to be aware of the facts that give an action its moral significance is to be aware of the morally relevant contents of the attitudes which guide that action (2014a: 102). Recall from Chapter 2 that for Levy, implicit biases "express nothing more than facts like: there is a statistical association between being male and being a police chief" (2014: 102) and so they have no morally relevant attitudinal content of which to be aware. So, one way to spell out the claim in NL1 is that, because there is no morally relevant attitudinal content of which to be aware in the case of actions guided by implicit bias, we are not morally responsible for implicitly biased actions.

But I think that this SD argument cannot be maintained, on two accounts. Firstly, I think that $\text{consciousness}_{PA}$ of the morally relevant facts will turn out to be an overly demanding condition for agents to meet in order to be morally responsible, and a condition according to which a lot of intuitively responsible agents will turn out not to be responsible. For instance, recall my example above of Katie, a person who harbours explicit prejudices against Muslims. Katie fails to acknowledge that she is biased in thinking that all Muslims are violent, and in doing so, fails to be conscious_{PA} of the facts that give her actions (such as her harmful speech acts during her interactions with Muslims) their moral significance. It is probable that many explicitly prejudiced agents are like this.

Explicit prejudice seems like the sort of thing for which we could and often are morally responsible. It seems appropriate for others to often have particular negative reactive attitudes to, for instance, explicitly prejudiced people who make harmful speech acts, even though these people may fail to acknowledge their biases.

Secondly, I think that there are at least some circumstances in which implicitly biased agents *are* aware of the content of an implicit bias *as* biased, as well as how this attitude may influence action. As I argued above, it is possible for our implicitly biased hirer Ben, as well as for Borgoni's 'Emilia' (2015), to be aware of their implicit biases *as* biased, and to catch themselves in an act mediated by such attitudes. Insofar as Ben and Emilia are (occurrently, introspectively) aware that their attitudes are biased, they are $\text{conscious}_{\text{PA}}$ of the morally relevant facts when those attitudes guide behaviour. So, once again, whilst this does not show that Ben or Emilia *are*, in fact, morally responsible for acting on their implicit biases, it does show that if they are excused, then it is not on the basis of their awareness of implicit bias and its influence on behaviour.

So, if it turns out that we do lack moral responsibility for all of our implicit biases, and their influence on our actions, then it will not be because we lack awareness of them, but because of some other distinguishing feature. I now turn to SD(R) claims for another such candidate set of distinguishing features: that of the structure of implicit biases and the way in which they are processed.

CHAPTER 4: RESPONDING TO SUBSTANTIAL DISTINCTION CLAIMS ON THE BASIS OF STRUCTURE AND PROCESSING

In the previous chapter, we saw that a fundamental distinction in kind between (i) implicit biases, and the actions that they guide; and (ii) agential attitudes such as beliefs, and the actions that they guide, could not be upheld on the basis of the kind of awareness that we have of each. This chapter examines arguments for the substantial distinction account of implicit bias on the basis of how the attitudes in question are structured and processed: and in particular whether they encode or respond to propositional information.

SD theorists such as Gendler and Levy suggest that implicit attitudes (implicit biases being among them) are to be distinguished from beliefs because only the latter are structured and processed propositionally. Contra Gendler and Levy, I argue that at least some of our implicit biases are propositional in structure, and feature in evidence-sensitive inferential transitions in the same way that many beliefs do. The conclusion of the chapter is that there is no substantial distinction between (i) the structure of implicit biases, and the way in which they are processed; and (ii) the structure of beliefs, as an example of agential attitudes, and the way in which they are processed.

In §4.1 I outline the psychological theory (dual process theory) which supposedly supports the claim that implicit biases are structured and processed associatively. According to dual process theory, there are two kinds of mental states, which are processed in fundamentally different ways: Propositional mental states encode relational information that holds between their constituent concepts, and are processed in accordance with their semantic content. Associative mental states encode nothing more than the frequency with which a person has experienced their constituent concepts together, where the activation of one concept primes the activation of an associated concept. According to dual process theory, this can happen regardless of the proposition implied by the activation, or of any propositional attitudes which the person has towards the constituent components. Some philosophers, such as Gendler in two papers published in 2008, have used this apparent distinction to argue that there is a fundamental difference in kind between beliefs and implicit biases.

I show that these philosophical claims generate two testable hypotheses: HYPOTHESIS 1 is that implicit biases are necessarily associative. HYPOTHESIS 2 is that beliefs change in response to changes in evidence. HYPOTHESIS 1 may be disproven by finding at least a few implicit biases which update in accordance with propositional information. I summarise findings from both de Houwer (2014) and Mandelbaum (forthcoming) to this effect in §4.2. I argue that whilst these findings survey relatively few implicit biases, even if just a few implicit biases are shown to be propositional, then this is sufficient motivation for rejecting the predictions of the dual process theory, that as yet untested implicit biases will be associative.

I discuss HYPOTHESIS 2, that beliefs change in response to changes in evidence, in §4.3. I argue that HYPOTHESIS 2 may also be shown to be false, because, as I demonstrate in §4.3.1, there are a number of examples of beliefs which fail to update in accordance with new evidence. However, I acknowledge that there is an ambiguity in the SD theorist's claims which generate HYPOTHESIS 2, and that these claims may have been intended to be interpreted normatively, rather than descriptively. I argue in §4.3.2, however, that the normative interpretation also fails to distinguish between implicit biases and beliefs, by showing that implicit attitudes may be governed by the same epistemic norms as beliefs.

Levy (2015) has recently accepted the falsity of HYPOTHESIS 1: he agrees that implicit biases may be propositionally structured. However, as I will demonstrate in §4.4, he has also provided a reinterpretation of HYPOTHESIS 2, which is supposed to reinstate the substantial distinction between implicit attitudes and beliefs. In particular, whilst he accepts that some implicit attitudes may be sensitive to propositional information, and so able to feature in inferential transitions, he argues that only beliefs are 'inferentially promiscuous', featuring in a much broader range of inferential transitions. I present this account in §4.4.1. In §4.4.2, I argue that if Levy's distinction between the inferential promiscuity of beliefs and the inferential sensitivity of at least some implicit attitudes is supposed to be a distinction in kind, then it will be at best an arbitrary one. If there is such a distinction in kind, then it must be the case that the most evidence sensitive implicit attitude is still evidence sensitive to a lesser degree than the least evidence sensitive belief. I demonstrate that, to the contrary, there are in fact some beliefs (in particular, *explicit* prejudices) which are evidence sensitive to a lower

degree than the most evidence sensitive implicit attitudes, and conclude that Levy's reinterpretation of HYPOTHESIS 2 does not support a substantial distinction between beliefs and implicit biases on the basis of their relation to evidence.

In light of the argument above, I then consider the SDR claim that the structure and processing of implicit biases rules out moral responsibility for implicitly biased attitudes and the actions that they influence. As in the conclusion to the previous chapter, I argue that this argument does not work: if it turns out that we do lack moral responsibility for our implicit biases, and their influence on our actions, it will not be because of their structure and the way in which they are processed.

4.1. THE EMPIRICAL BACKGROUND TO THE NOTION OF ASSOCIATIONS

In the following, I will present a brief overview of a position in cognitive science known as 'dual process theory', and demonstrate how the SD arguments on the basis of structure and processing rely on claims made by dual process theorists.

4.1.1. Dual process theory and single-process theory

The notion that implicit bias implicates associative processes originates in cognitive science. Some psychologists, such as Sloman (1996) posit that the mind is comprised of an 'Associative System' and a 'Rule Based System' which process mental entities in two distinct ways: the rule based system processes mental entities in virtue of their propositional contents, whereas the associative system processes mental entities just in virtue of how closely they are associated with each other—in a sense that I will explore in more depth shortly. This 'dual processes' interpretation of a range of findings in cognitive science is also favoured by Strack and Deutsche (2004) as well as Gawronski and Bodenhausen (2006, 2011, 2014).

Three things are important to note as regards the dual process interpretation, and my argument to follow. Firstly, the dual process interpretation is not incontrovertibly implied by the psychological evidence. Rather, proponents argue that the dual process interpretation *best explains* the findings, and *best predicts* how people will perform in experimental conditions, (for example see Gawronski & Bodenhausen, 2014).

Secondly, there is no consensus across the empirical community that the dual process interpretation is correct. A competing hypothesis, the 'single-

process' interpretation, is favoured by other cognitive scientists such as Fazio (1990), Olson & Fazio (2008), Petty & Briñol (2006), and Petty *et al.* (2007). For single-process theorists, there is no distinction between associative processes and rule-based processes. Instead, the mind is comprised of just one system which processes all mental entities in fundamentally the same way.

Thirdly, it is beyond my remit in this thesis to provide an argument for whether the single-process model or the dual process model is, in general, the correct interpretation of mental processing. But such a verdict is not necessary for my purposes because my argument will be that implicit biases are not necessarily associatively structured states, and that at least some implicit biases implicate rule-based processes, rather than associative processes. It is consistent with the data and argument presented in this chapter that the dual process model could still hold in general, in that some set of mental entities (other than implicit social attitudes) may be processed by an associative system, which is distinct from a rule-based system. So, my argument against the SD theorist that there is not a substantial distinction between implicit biases and beliefs on the basis of structure, content or processing is independent of the debate over whether the dual process or single-process model is the correct interpretation of mental processing in general.[33]

Let us now turn to the supposed distinction that dual process theorists maintain exists between the rule-based and the associative system, before assessing whether implicit biases are in fact associatively structured and associatively processed states. Let's look first at the rule-based system. Many mental states, such as beliefs or desires, for instance, contain constituent concepts. For example, the belief that 'trees are green', call it *B1*, contains the concept of TREES and the concept of the colour GREEN; and the belief, *B2,* that 'Dan loves George' contains the constituent concepts DAN and GEORGE. Further, at least some mental entities specify the particular way in which their constituent concepts

[33] Whilst my argument against the SD theorist about implicit biases is independent of the general dual process *vs*. single-process debate, the reverse may not be true, because dual process theorists often appeal to at least some of the data on implicit bias to justify why they think that the dual process interpretation is correct, (for example, see Gawronski & Bodenhausen, 2014: 450). If appeal to this data on implicit bias plays a necessary role in the argument for why dual process theory is (supposedly) a superior interpretation to single-process theory, then the arguments in this chapter that at least some implicit biases are not associatively structured or associatively processed might well put some pressure on at least some arguments for the dual process theory. But as I explained in the main text, I do not have the scope to pursue this further in the thesis.

are related—the *kind* of relation that holds between them. This is the case for both *B1* and *B2*. In *B1* the colour green is a property that is had by trees, and in *B2* Dan and George are related through Dan's love of George.

Philosophers term these sorts of mental states 'propositional attitudes,' and because these attitudes represent propositions, they have a semantic content which is truth evaluable. The specification of how the constituents of the proposition are related at least in part determines the semantic content of the attitude in question. For example, 'Dan loves George' has a different semantic content to 'George loves Dan', even though both propositions have the same constituents: Dan, George, and the relation of 'love'. Attitudes with semantic content may enter into what Levy calls 'inferential transitions' in accordance with the semantic content in question (2015). That's to say that they may imply or be implied by other propositions according to the semantic content specified by the constituent concepts and the relation that holds between them. For instance, a person who believes *B1* as above, and who is told that 'there is a tree outside', may infer that 'the tree outside is green'; whilst someone who believes *B2* as above, and who is told that 'Dan has bought a lovely birthday present for the only person he loves', may infer that 'Dan has bought a lovely birthday present for George'.

We can see that in order for a mental system to process propositional attitudes and produce the kind of inferentially valid transitions discussed above, that system must be sensitive to the semantic contents of a propositional attitude—to the particular way in which the constituent concepts of the propositional attitude are related. So, at least some of the time, it seems that we process mental states in virtue of a system that is sensitive to semantic relations, and which can produce valid inferential transitions. For dual process theorists, this kind of inferential, rule-based processing is the preserve of the rule-based system.

In setting up the case for the existence of a second, distinct system, dual process theorists rely on evidence which shows that the activation of one concept primes a set of other concepts which are then apt to feature in processing or to affect behaviour. The set which gets primed are those concepts, examples of which the person in question has previously experienced as "spatiotemporally contiguous" with the first concept in their environment (Gawronski & Bodenhausen, 2014: 453)—or, in other words, entities that the person in question has previously encountered alongside an entity instantiating the first concept. Recall from Chapter 1, for instance, Meyer and Schvaneveldt's (1971) finding

that participants are quicker to recognise 'butter' as an English word when first presented with the word 'bread', compared to when they were first presented with a word that is unrelated to 'butter' (such as 'window' or 'doctor'). The dual process interpretation of these results is that bread and butter are experienced as 'spatiotemporally contiguous', or seen together, more often than butter and windows are seen together, and so subjects develop a strong associative link concerning those concepts. So, when one concept is activated, concepts which have a strong associative link with the first concept become primed for activation, making them more easily accessible to mental processing, and apt to affect behaviour, than concepts which are less closely associated with the first concept.

Dual process theorists hold that the priming of closely associated concepts can occur even if the person in question has propositional attitudes which imply that the associated concepts are in fact *in*appropriate to the particular context in question, (Gawronski & Bodenhausen, 2014: 450).[34] For example, for someone who is asked "What is your favourite colour?" it may be the case that the word 'colour' activates the concepts of RED, ORANGE, GREEN, etc., even when the person in question rejects at least some of the activated concepts as aesthetically pleasing colours.

It should be noted that it is not the case, for dual process theorists, that when the activated concept A primes the associated concepts B, C and D, that B, C and D will *inevitably* feature in subsequent mental processing or action guidance. Dual process theorists hold that, at least sometimes, it is possible for rule-based processing to override the influence of concepts which are primed by the associative system. Instead, the commitments of dual process theory are that (i) after A becomes activated, B, C and D are *more likely* to affect behaviour than concepts E, F and G where E, F and G are less closely associated with A than B, C and D; and (ii) that after A becomes activated, B, C and D will be more easily accessible, and so they will affect behaviour more quickly than E, F and G in

---

[34] Gawronski and Bodenhausen argue that "mentally associated concepts can be activated regardless of whether the relation implied by the activated link is considered valid or invalid," (2014: 450). They support this by appealing to findings of implicit bias, saying "For example, encountering a Muslim-looking man may activate the concept *terrorism* even if a person rejects the implied connection between Muslims and terrorism," (Gawronski & Bodenhausen, 2014: 450). As per the previous footnote, this is just one example of where dual process theorists rely on implicit biases as paradigm examples of entities which are associatively processed. I'm avoiding giving an example based on implicit bias in the exposition of the notion of the associative system, because whether or not implicit bias really *is* associative in this way is exactly what is at issue in this chapter.

circumstances where the rule-based system cannot operate, such as when responding as quickly as possible to an experimental sorting task like the IAT. According to dual process theorists, the rule based system requires more time to process concepts than the associative system requires to prime them for use in further processing and behaviour, and so when people must act quickly, such as when they are pressed to respond to various experimental tasks, the rule-based system does not have time to override the operation of the associative system.

A brief clarification: Up until this chapter, the thesis has been mainly concerned with implicit biases conceived of as mental states, and their propensity to influence actions. This section introduces the notion of the systems which *process* these mental states. This might appear to open up the conceptual possibility that an associative mental state could be processed by the rule based system, whilst a propositional mental state could be processed by the associative system. But this is not what dual process theorists think is the case. In fact, as we can see from the passage below, associative processes and associative states are somewhat inter-defined, and likewise for propositional processes and propositional states:

> A central assumption of [the main dual process model]…is that implicit evaluations reflect the behavioral outcome of *associative processes*, whereas explicit evaluations are the behavioral outcome of *propositional processes*. Associative processes are defined as the *activation* of mental associations in memory, which we assume to be driven by the principles of feature matching and spatiotemporal contiguity. Propositional processes are defined as the *validation* of the information implied by activated associations, which we assume to be guided by the principles of cognitive consistency. (Gawronski & Bodenhausen, 2014: 449)

So, according to dual process theorists, associative processes are those which process associations on the basis of their spatiotemporal contiguities, and propositional processes are those which assess the semantic content of activated mental states, accepting such content if it is considered to be consistent with other propositional attitudes, or rejecting it if it is considered to be inconsistent.

4.1.2. Conditions of acquisition, modulation and extinction

For dual process theorists, there is a fundamental distinction between the way in which associations are originally encoded and the way in which propositions may be encoded. According to dual process theory, associations may only be formed "gradually as the result of many experiences" (De Houwer 2014: 343). In order for the associative system to produce an association between concept A and concept B, a person must witness the co-occurrence of examples of concept A alongside examples of concept B many times over. Dual process theorists Gawronski and Bodenhausen hold that:

> The central assumption underlying this definition [of associative processes] is that observed co-occurrences between objects and events result in a co-activation of their corresponding mental concepts, which in turn creates an associative link between the two. Repeatedly observing the same co-occurrences strengthens this link, which facilitates the spread of activation from one concept to the other upon encountering one of the two associated stimuli. (2014: 453)

Propositions, on the other hand, can be formed "as the result of a single instruction or inference" (De Houwer 2014: 344). For dual process theorists, this process is supposed to be distinctive to the formation of propositions alone. For instance, the proposition TREES ARE GREEN might be formed simply by being informed by a reliable person that "trees are green." Gawronski and Bodenhausen suggest that when people are presented with new propositional information, they perform a "validity assessment" of that information, wherein it "may be regarded as either true or false depending on its consistency with other momentarily considered propositions," (2014: 453), and they form a new proposition accordingly.

Both de Houwer (2014: 344) and Gawronski and Bodenhausen (2014: 453) agree that the dual process theory allows that propositions may also be formed as a result of seeing the co-occurrence of two objects multiple times. However, Gawronski and Bodenhausen claim that in order to form a new proposition in this manner, the considered information must still "pass a process of propositional validation" which "involves the acquisition of self-generated propositional information," (by which I think they mean that a person *infers* a new proposition from their previous experiences of the co-occurrence of two

concepts). For instance, a person might see multiple instances of trees with green leaves, and infer, on the basis of these experiences, the proposition TREES ARE GREEN. According to Gawronski and Bodenhausen, this process of inference does not happen in the case of the formation of an association, and so for them, even if not necessarily for de Houwer, the process of proposition formation is importantly distinct from association formation.

Alongside the claim about how associations are formed, dual process theorists are committed to a symmetrical hypothesis about how associations are modified. According to Gawronski and Bodenhausen, "repeated co-occurrences in the environment may create new associative links between concepts in memory," (2014: 454). For them, the only way to alter an already encoded association between concepts A and B is through a process of 'extinction' or 'counter-conditioning': Either A is presented many times over with ¬B, (extinction), or, if B is not the kind of thing that can be negated, then with an alternative concept C (counter-conditioning). Even then, this does not guarantee that an existing association *will* be changed, or extinguished (Gawronski & Bodenhausen, 2014: 454).

Importantly, for dual process theorists, a single propositional instruction or inference should not modify an association. Indeed, according to Gawronski and Bodenhausen, focusing on propositional information that negates the implied content of an association "leaves the activation of associations unaffected" or even "produce[s] ironic effects", whereby the association is activated to affect behaviour more significantly than if the person in question had not focused on the negating propositional information, (2014: 455). The claim that a single propositional instruction or inference should not modify an association will be important when it comes to testing the SD theorist's claims, as I outline in the next subsection.

So, to summarise, dual process theorists hold that the mind is comprised of two systems which process information in two distinct ways. The rule-based system processes propositional mental entities (such as beliefs) in accordance with their semantic content and inferential relations. The associative system primes concepts for use in further processing and behaviour in virtue of how closely they are associated to one another in memory (which is determined by how often the person has witnessed their co-occurrence), regardless of whether the person in question endorses the implied content of an association or not.

4.1.3. Substantial Distinction views, and some testable hypotheses

In light of the above, we're now in a position to appreciate the empirical background to the SD claims introduced in Chapter 2, in which implicit bias is said to implicate associative mental states and processes:

NL5: Implicit biases are associative, not propositional, in structure.[35]

TG1: Implicit biases are aliefs: *sui generis* tripartite mental states with a representational component, an affective component, and a behavioural component, which are 'associatively linked'.

TG3: Implicit biases, insofar as they are aliefs, are not sensitive to the propositional information encoded in mental states such as beliefs: Learning that not-*P* may well not cause me to cease having an implicit bias with the apparent content that *P*.

Indeed, Levy's claim summarised in NL5 comes from a book in which he gives a substantial exposition to the above kinds of empirical considerations, in much more depth than I am able to here.[36] Gendler's claims about the associative nature of implicit biases apply insofar as she thinks that implicit biases are examples of *aliefs*, a *sui generis* class of mental state. Importantly for her, the representational, affective and behavioural components of an implicit bias are not linked in virtue of a relation which bears semantic content, but in virtue of associations between the representational component, the affective response, and the behavioural output. These components tend to be co-activated, and are not sensitive to the propositional content of more familiar attitudes such as beliefs (Gendler, 2008a: 651).

The general claim, common to both Gendler's and Levy's substantial distinction positions that implicit biases are structured and processed associatively, delivers a testable hypothesis:

HYPOTHESIS 1: implicit biases are necessarily associative

[35] This view is from Levy's 2014a book, and it is a view that he rejects in his 2015 paper, as I will discuss in §4.4.

[36] See chapter 3 of Levy's 2014*a*.

This hypothesis may be shown to be false if at least one implicit bias which updates in accordance with propositional information can be found: as we just saw, according to the psychological theory on which these claims are based, if implicit biases are associative, then they should not update in accordance with propositional information, (this point is also emphasised by de Houwer, 2012 and Mandelbaum, forthcoming).

In addition to Gendler's characterisation of aliefs (the class to which she says implicit biases belong) she makes some further claims regarding the characteristics which are particular to *beliefs,* though as noted in Chapter 2, it was unclear whether these claims were to be understood as descriptive or normative claims.

Here, the claims look descriptive:

> If I believe that *P*, and subsequently learn that not-*P*, I will revise my belief... Learning that not-*P* may well not cause me to cease alieving that *P*... alief just is not reality-sensitive in the way belief is. Its content does not track (one's considered impression of) the world. (2008a: 651)[37]

> Beliefs change in response to changes in evidence; aliefs change in response to changes in habit. If new evidence won't cause you to change your behaviour in response to an apparent stimulus, then your reaction is due to alief rather than belief. (Gendler, 2008b: 566)[38]

On a descriptive interpretation, the claim is that beliefs *always* update in accordance with the proposition implied by new evidence. This can be seen in the latter sentence of Gendler's second quote above, which implies that it is a necessary condition of any belief that both it, and the behaviours it brings about, update in accordance with new evidence.[39] If the claim about this particular characteristic of beliefs is descriptive, then it is also testable in the following way:

---

[37] Note that this is the basis of what I called claim TG3.
[38] Note that this is the basis of what I called claim TG4.
[39] I take it that, on the descriptive interpretation, Gendler needs it to be the case that beliefs necessarily update in light of evidence, rather than something weaker, such as that beliefs only typically update in light of evidence, in order to support her (SD) claim that aliefs are a fundamentally different kind of mental state to beliefs.

> HYPOTHESIS 2: Beliefs necessarily change in response to changes in evidence. If I believe that *P*, and subsequently learn that not-*P*, I will revise my belief that *P* to a belief that not-*P*.

To show that HYPOTHESIS 2 is false, we need to find at least one belief that not-*P* which fails to be revised in response to learning that not-*P*.
Elsewhere, however, Gendler seems to make a normative claim when she argues that beliefs update in accordance with evidence, saying:

> belief aims to 'track truth' in the sense that belief is subject to immediate revision in the face of changes in our all-things-considered evidence. When we gain new all-things-considered evidence—either as the result of a change in our evidential relation to the world, or as a result of a change in the (wider) world itself—the norms of belief require that our beliefs change accordingly. (Gendler, 2008b: 565)

The talk of 'aims' and the notion of being 'subject to' immediate revision, as opposed to *necessarily* being immediately revised in the face of new evidence suggests that beliefs are distinguished from implicit biases (insofar as they are supposed to be aliefs) because the former are governed by some set of norms which require, but do not guarantee, that they will update in accordance with new evidence. If the normative interpretation of the claim is correct, then it may not be refuted simply by finding examples of beliefs which do not in fact update in accordance with evidence. This result is still consistent with the normative interpretation, as long as the beliefs in question are governed by the relevant evidence update norms. Refuting the normative interpretation, then, is a matter of either (a) rejecting that there is such a thing as an evidence norm, or (b) arguing that whatever the evidence-sensitivity norm turns out to be, it could *also* govern implicit attitudes. (I will argue for (b).)

My intention over the next two sections is to show that both HYPOTHESIS 1 and HYPOTHESIS 2 are false. I address HYPOTHESIS 1 in §4.2, where I give several examples of implicit biases and other implicit social attitudes which *do* update in accordance with evidence. Following this, HYPOTHESIS 1 is false, and any substantial distinction claims which rely on the notion that implicit biases are necessarily associative in structure, and in the manner in which they are processed, will fail. In §4.3 I consider Gendler's evidence update claims as

regards beliefs. I address the descriptive version (HYPOTHESIS 2) in §4.3.1. There, I give several examples of beliefs which fail to change in response to changes in the agent's evidence, thus showing HYPOTHESIS 2 to be false. I then turn to the normative interpretation of Gendler's evidence update claim in §4.3.2, where I argue that evidence-sensitivity norms may *also* govern implicit attitudes, and hence, implicit biases. So, on both the descriptive interpretation, and the normative interpretation of Gendler's claims, the substantial distinction argument fails: there is no substantial distinction between beliefs and implicit biases on the basis that the former either do, or should, update in light of evidence whilst the latter do not.

4.2. HYPOTHESIS 1 FAILS: IMPLICIT BIASES ARE NOT NECESSARILY ASSOCIATIVE

As we saw above, according to dual process theorists, implicit attitudes are paradigm associative states (Gawronski & Bodenhausen, 2014: 449). Recall also that, for dual process theorists, the only way to alter an association between concepts A and B is through a process of 'extinction' or 'counter-conditioning': Either A is presented many times over with ¬B, (extinction), or, if B is not the kind of thing that can be negated, then with an alternative concept C (counter-conditioning), (Gawronski & Bodenhausen, 2014: 454-5).

So dual process theorists are committed to the claim that if a mental state can be altered through processes other than extinction or counter-conditioning, then the mental state in question is *not* an associative one, (Mandelbaum, forthcoming: 17). SD theorists who rely on dual process theory to maintain that implicit biases are fundamentally distinct from beliefs because the former are associative states whilst the latter are propositional states, are thereby committed to the claim that no implicit biases may be altered through processes other than extinction or counter-conditioning.

Problematically for these SD theorists, there is evidence to show that a number of implicit social attitudes, including the subset that we're interested in, implicit biases, may be modulated by processes other than extinction or counter-conditioning. Both de Houwer (2014) and Mandelbaum (forthcoming) discuss a number of relevant findings, which I outline below.

Implicit biases are modulated by strength of argument

Strength of argument can affect the strength of an implicit bias. Briñol et al. (2008, cited in Mandelbaum, forthcoming) demonstrate that subjects who are presented with a strong argument for hiring an African American professor (citing their academic merits) exhibit less bias on a subsequent IAT than that exhibited by those presented with a weak argument (citing the benefit to the image of the institution). According to the dual process interpretation on which implicit biases are associative, just the mention of the term 'African American' should activate negative implicit associations which could then be observed on a subsequent IAT. Both the strong and the weak argument contained the same number of mentions of the term 'African American professors'. So if associations mediate action on the IAT, then we would expect there to be no significant difference between the performance of the two groups. But this was not what was observed—those who read a strong argument for hiring African American professors exhibited less anti-African American bias on the IAT. All else being controlled for, Briñol et al. (2008) conclude that strength of argument is the variable that accounts for the resulting difference in the level of implicit bias across conditions.[40]

Implicit social preferences are modulated by relational information

'The enemy of my enemy is my friend' goes the old adage. This saying exemplifies the basic semantic content of 'enemy' and 'friend' as logically complimentary relations, and is predictive of explicit preferences, (for example, see Heider, 1958; Aronson & Cope, 1968; as referenced in Mandelbaum, forthcoming). According to dual process theory, associative processes are blind to the propositional notion of double-negation elimination: a negated negative valence is processed as a negative valence. So, if implicit like/dislike preferences are associative, then we would expect an enemy's enemy to inherit a negative valence, on account of being associated with two negative items.

However, implicit preferences which are sensitive to semantic content of enemy-friend relationships have been observed. Gawronski *et al*. (2005, cited in Mandelbaum, forthcoming) presented subjects with a series of photos of unfamiliar people (the 'CS1s') which were coupled consistently with either

[40] Note that this result refutes Levy's claim in NL7: We can influence our implicit attitudes only indirectly, by attempting to form new associations.

positively or negatively valenced concepts. Experimenters then presented subjects with a second series of photos of different people (the 'CS2s'), as well as information on whether the CS2s were liked or disliked by the CS1s. As Mandelbaum points out, the associative theory predicts that:

> …you should have enhanced negative reactions toward the CS2 because you a) are encountering the CS2 as yoked to negative CS1 and b) are activating another negative valence because you are told that the CS1 dislikes the CS2. (forthcoming, 11)

However, the results showed quite the opposite. Participants exhibited implicit *preferences* for the CS2s who were disliked by negatively valenced CS1s. That is, participants exhibited implicit preferences for the enemies of their enemies. These results imply that implicit preferences are sensitive to the logical notion that the negation of a negative property is equivalent to a positive property—sensitive to propositional information. Such results cannot be explained by a theory on which implicit like/dislike preferences are associative, (Mandelbaum, forthcoming: 11).

Peters and Gawronski (2011) conducted a later study in a similar paradigm, (in DeHouwer, 2014). They presented participants with pictures or four unknown people alongside a series of personality traits. Persons A and B were mostly presented alongside positive traits, whilst persons C and D were most often presented alongside negative traits. Participants were informed that persons A and C were paired with words which truly described them, whilst B and D were paired with words that are the opposite of their actual traits. Participants' evaluations of the four people were then tested, testing which including an IAT test of the implicit attitudes participants held towards A, B, C and D. If these attitudes were formed associatively, purely on the basis of seeing the people presented in spatiotemporal contiguity with the relevant traits, then we would expect that persons A and B (who were mostly shown alongside positive traits) would be evaluated more positively than C and D (shown alongside negative traits). Even though person A was evaluated more positively than person C (which is predicted on the associative hypothesis), it turned out that person B was evaluated *less positively* than person D. It appears that the propositional content of the relational information supplied at the start, that persons B and D had the *opposing* traits to those they would be shown alongside, had a modulating effect on the participants'

implicit evaluations. This difference in evaluation cannot be explained by the associative hypothesis: An associative hypothesis predicts learning that B and D were shown alongside *opposing* traits to those which they in fact had will have no difference on implicit evaluations.

The results from Gawronski *et al.* (2005, in Mandelbaum, forthcoming) and Peters and Gawronski (2011, in DeHouwer, 2014) reveal that at least some implicit social preferences and evaluations are modulated by propositional information, leaving the associative theory of implicit evaluations unable to explain these results.

Implicit attitudes about (fictional) social groups are modulated by propositional instruction

New implicit attitudes, which are formed as the result of an abstract propositional instruction, can be as virulent as new implicit attitudes formed as the result of extensive associative conditioning. Perhaps even more surprisingly, further propositional information modulates these implicit attitudes to a greater extent than further associative conditioning.

In a study by Gregg et al. (2006; in Mandelbaum, forthcoming), half of the subjects read a single sentence about which of two (fictitious) tribes are peaceful and civilised, and which are savage and barbaric. The other half underwent 240 trials to associatively condition one tribe with the notions of peace and civilisation, and the other with savagery and barbarism. (The same tribes were matched with the same notions across conditions). Both groups were then given an IAT test. Those in the propositional instruction condition showed implicit attitudes coupling tribes with concepts which were as virulent as those in the associative learning condition. In other words, 240 trials of associative conditioning resulted in the association of two terms no more significantly than attitudes formed as the result of a single propositional instruction. Being presented with an abstract proposition in which a social group are predicated with a valenced concept produced *as efficacious* a behavioural response on the IAT as extensive associative conditioning. This is problematic for dual process theorists, who hold that the strength of an association (as revealed on the IAT) correlates with how often concepts appear in spatiotemporal contiguity, (Gawronski & Bodenhausen, 2014). That a group who have seen two concepts in spatiotemporal contiguity once have as strong an implicit attitude as those who have seen the

concepts in spatiotemporal contiguity 240 times is not explainable on the associative hypothesis (Mandelbaum, forthcoming: 16).

A later study in the Gregg et al. (2006) is perhaps even more problematic for the associative hypothesis. Participants in the propositional instruction condition were instructed to read a sentence which contradicted the contents of the first sentence, i.e. a sentence in which the adjectives which originally described the first tribe now describe the second, and vice versa. Those in the associative conditioning trial underwent extensive counter-conditioning trials in which the adjectives which were originally presented alongside the first tribe were now presented alongside the second, and vice versa. This counter-conditioning failed to have any effect on subjects' attitudes. An IAT in the counter-conditions group revealed attitudes in accordance with the originally conditioned valences which were as efficacious as before the counter-conditioning. However, an IAT of those in the propositional instruction condition revealed that their implicit attitudes *had* in part adjusted in line with the new information: exhibiting implicit attitudes which were less extreme than their attitudes on the first test, and less extreme than those in the associative condition. Their implicit attitude did not reverse completely as the result of a single propositional instruction. However, as compared with those in the associative conditioning group, propositional instruction was revealed to be *more effective* at modulating implicit attitudes than associative counter-conditioning. The associative hypothesis cannot make sense of this, as Gregg *et al.* acknowledge in the following:

> Our first two experiments therefore empirically contradict what dual process models can plausibly be taken as implying, namely, that automatic attitudes are relatively immune to sophisticated symbolic cognition (2006: 9; quoted in Mandelbaum, forthcoming).

DeHouwer (2006) shows that the modulation of implicit attitudes by propositional information also occurs in non-social attitudes. Study participants were given some instructions which inform them that that they will be presented with a series of pleasant and unpleasant pictures, where the pleasant pictures will always be preceded by the brief presentation of two neutral words, and the unpleasant pictures preceded by two different neutral words. Before the presentation of pictures began, subjects underwent an IAT which revealed that they were

significantly faster at pairing the neutral words that they were told were going to accompany the pleasant pictures with pleasant evaluative concepts, and to pair the neutral words that they were told were going to accompany the unpleasant pictures with unpleasant evaluative concepts than the other way around (i.e. pairing neutral words in a manner incongruent with the instructions). DeHouwer suggests that "a single instruction about the relation between the meaningless words and positive or negative pictures was sufficient to influence the implicit evaluation of the words" (2014: 347).

In light of the above evidence, it appears that implicit bias, as well as implicit evaluation more generally, may be modulated by the strength of an argument, by a single propositional instruction, and by relational information. None of these results are explainable on an associative hypothesis, but they are both explainable and predicted by a hypothesis on which the implicit attitudes in question both encode propositional information, and are sensitive to its content. These findings undermine SD views according to which there is supposed to be a fundamental distinction between all implicit biases and all beliefs on the basis that the former are associative in structure, whilst the latter are propositional in structure. Recall Levy's (2014a) claim, as summarised in NL5, and Gendler's (2008a, 2008b) claim, as summarised in TG1 and TG3:

NL5: Implicit biases are associative, not propositional, in structure.

TG1: Implicit biases are aliefs: *sui generis* tripartite mental states with a representational component, an affective component, and a behavioural component, which are 'associatively linked'.

TG3: Implicit biases, insofar as they are aliefs, are not sensitive to the propositional information encoded in mental states such as beliefs: Learning that not-*P* may well not cause me to cease having an implicit bias with the apparent content that *P*.

Importantly, for both Levy and Gendler, implicit biases are characterised by their being associatively structured, and not responsive to propositional information. This generated HYPOTHESIS 1, that implicit biases are *necessarily* associative in structure, and so necessarily processed associatively. So, even if just one result shows that implicit biases may be modulated by propositional information, then the hypothesis that implicit biases are fundamentally distinct from beliefs because

they are structured and processed associatively, whilst the latter are structured propositionally, and processed in accordance with propositional information, fails.

It might be pointed out that whilst these examples show that at least some implicit social attitudes are processed in accordance with propositional information, and so are structured propositionally, it doesn't show that all implicit social attitudes are so structured and processed. It might then be argued that, for all that we have shown above, the majority of as yet untested implicit social attitudes are not propositional, but associative.

I don't think that this is quite right. Recall from §4.1 that the dual process theory, the idea that there is an associative system which processes associations in a distinct manner to that of the propositional system, is a *theoretical model*. It is not incontrovertibly implied by the psychological findings. Rather, it is upheld insofar as it serves to both explain and to predict the findings (Gawronski & Bodenhausen, 2014). The findings above are neither explainable, nor predictable on a dual process hypothesis, according to which implicit attitudes, by their very nature, are associative.[41] The dual-systems model only has theoretical utility, with respect to implicit social attitudes at least, if *all* implicit social attitudes are associative. If some implicit social attitudes are in fact propositional, then, alongside the dual-systems model, we need to commit to another model to account for those attitudes which the dual systems model cannot explain. But if we have to commit to another model, then we lose the motivation for positing the dual systems model in the first place. So, even though we haven't yet empirically verified that all implicit biases are indeed propositional, we have lost a considerable theoretical reason to think that they might be otherwise. It may well be that the dual process model has theoretical utility independently of the debate regarding implicit social attitudes, over territory where the competing single systems models cannot account for the data, and so I do not take this to be an outright refutation of the duals systems model in general. However, given that the dual process model cannot explain the propositional structure of at least some implicit social attitudes, we have little reason to place faith in its predictions of the structure of as yet untested implicit social attitudes.

---

[41] It is not an option for the dual processes theorist to deal with the evidence presented in this section by saying that that those implicit attitudes that exhibit propositional features turn out to be *explicit* attitudes after all. These implicit attitudes have other features which, according to dual process theorists, propositional processes should lack. For instance, they are non-effortful; they operate automatically; and do not require guidance from attentional resources.

4.3. HYPOTHESIS 2 FAILS ON BOTH A DESCRIPTIVE AND NORMATIVE INTERPRETATION

HYPOTHESIS 2 is the claim that beliefs necessarily change in response to changes in evidence. On a descriptive interpretation of this claim, finding just one belief which fails to change in response to changes in the agent's evidence is sufficient to show that it is false. On a normative interpretation, the claim fails if both beliefs and implicit attitudes are governed by evidence sensitivity norms. In what follows, I will show that both the descriptive interpretation and the normative interpretation of the evidence claim fail to uphold a substantial distinction between implicit biases and beliefs. I start with the descriptive interpretation.

4.3.1. The descriptive interpretation fails

The descriptive reading of Gendler's (2008a, 2008b) claims that beliefs update in light of evidence, which I summarised as key claims G3 and G4, generates HYPOTHESIS 2, as follows:

> HYPOTHESIS 2: If I believe that *P*, and subsequently learn that not-*P*, I will revise my belief that *P* to a belief that not-*P*.

To show that HYPOTHESIS 2 is false, we must find at least one belief that *P* which fails to be revised in response to learning that (which Gendler uses interchangeably with having evidence for) not-*P*.

One might think that any false belief is a belief which has not updated in light of the evidence. If this were how Gendler (2008a, 2008b) intended the claim to be interpreted, then her substantial distinction argument would fail for quite trivial reasons—plenty of beliefs are false. In light of this, I take Gendler to mean that an agent will update their belief in light of evidence which they recognise and interpret as relevant to that belief. So, the claim is not that all beliefs update in light of evidence, but that if an agent encounters evidence which they see as justification for the formation of a new (or alteration of an existing) belief, then they should form a new belief accordingly.

However, such a claim is just too strong for many ordinary cases that we still might want to countenance as cases of belief. Imagine that I believe (falsely) that you live on College Road, a belief I have held for a long time on the basis of

misunderstanding something you said in conversation years ago. Let's imagine that this comes up in conversation between us later on, whereupon you correct me—in that conversation from years ago you were in fact talking about helping your brother move in to College Road, rather than moving in yourself. You in fact live on Station Road. You're a generally reliable person, who rarely intends to deceive and so I see your testimony as justification for the formation of a new belief, that you in fact live on Station Road. Nevertheless, a few weeks later, when someone asks me where you live, I say "College Road", fully believing myself to have stated something true—our recent conversation about where you in fact live having slipped my mind. In this case, I've retained an old belief, in spite of having learned evidence to the contrary. So, cases of memory lapse are problematic for the notion that agents always update their beliefs in accordance with evidence which they see as justification for the formation of a new belief.

It might be objected that if I've forgotten that you told me you live on Station Road, then it's not true to say that my belief is unresponsive to the evidence that you live on Station Road, and so it is not a problem for the descriptive interpretation of Gendler's view. It might then be that beliefs only have to update in accordance with evidence that the subject, in the moment that the belief is tokened, sees as evidence. What is important, then, is that if a subject sees some evidence *E as* evidence for the proposition *P*, then in that moment, she will adopt the belief that *P* accordingly. But then, consider the following cases:

> *Self-doubt*
> Ada sees the fact that she consistently gets As and A*s on her maths tests as evidence that she is good at maths, and yet she does not believe that she is good at maths because she has crippling self-doubt about her own abilities.

> *Grief*
> Hakeem goes to identify the body of his brother, who died after being hit by a car. The doctor who tried to save his brother's life recounts what happened. From the doctor's account, it is evident that his brother was conscious, confused, and in much pain as the ambulance arrived, and he was attended to by paramedics. However, even in the face of this evidence, Hakeem cannot accept that his brother suffered, and finds himself believing that his brother died instantly.

Explicit prejudice

Mike is a neuroscience student who believes that female brains lack some of the processing capacities of male brains, and that this shows that women are cognitively inferior to men. He sincerely assents to this belief, asserts it frequently, and utilises it in judgements. Mike reads Joel et al.'s (2015) study which shows that human brains are not distinguishable on the basis of sex. He has sufficient training to understand the methodology and the conclusions. However, he fails to update his belief that women are inferior to men on the basis of the neurological evidence, and goes on believing that neuroscience shows that women are academically inferior to men.

Here, we have three agents who see the evidence that *P as* evidence that *P*, and yet fail to adopt the belief that *P*.

Let's take one of these agents, and look more closely at what is going on in their psychological economy. Let's take Ada. For the duration of Ada's presentation on trigonometry, Luke and Steven sit at the back of the class, snickering. For the next couple of weeks, Luke and Steven tease Ada in each maths class. Ada interprets this behaviour as evidence that she is bad at maths, and, because of this, forms the *belief* that she is bad at maths. Of course, Luke and Steven's actions aren't really evidence (or, at least, they are *bad* evidence) for the proposition that Ada is bad at maths. So Ada has formed a false belief. So far, this is consistent with Gendler's claims. Ada has ended up with a false belief, but only because she took herself to have evidence for the belief in question, and, for Gendler, taking oneself to have evidence that *P* is sufficient for believing that *P*.

Some time later, Ada receives her class report, which details all of the marks that she got in recent tests and exercises. The class average wavers around 60%, but Ada consistently achieves over 70%, with some scores in the 90s. Ada understands that this is good evidence that she is good at maths—and certainly better than the majority of her peers. However, her self-doubt is such that she cannot bring herself to believe that she is good at maths. She cannot assent to the proposition that she is good at maths. When she thinks about the subjects that she needs to work on most for the upcoming exams, she concludes that it is the numerate subjects that she is worst at. She thinks about possible future careers and resolves not to pursue any positions which require numeracy. She dreads the upcoming maths exam more than any of the others. In terms of her phenomenology, her reasoning and her actions, the proposition "I am bad at

maths" is playing a somewhat prolific role. The proposition "I am good at maths" plays no such role. On any account of belief, it looks as if Ada has *not* acquired the belief that she is good at maths. And yet she has been presented with something that she recognises as evidence for that belief.

I can imagine the following objection: If Ada does not end up with the belief that she is good at maths, then perhaps she does not really *fully* accept that she has evidence for this proposition in the first place. If that's the case, then the example is simply wrongly described. Ada recognises that she has some good marks, but she doesn't really accept this as decisive evidence for her being good at maths. The trouble with this line of argument, however, is that it is not clear what taking something to be evidence amounts to, if Ada doesn't do this with respect to her test scores. What entitles us to say that an agent has not really accepted the evidence that entails *P as* evidence that entails *P*? It had better not be an appeal to the fact that they have *failed* to adopt the entailed belief: If the argument that Ada has failed to fully appreciate the evidence amounts to nothing more than drawing attention to the fact that she has failed to form the appropriate belief, then we have a problem of triviality. The claim from HYPOTHESIS 2 that needs to be defended is that:

> If *S* recognises that there is evidence that not-*P,* then they will revise their belief that *P* (to a belief that not-*P*).

If the means of defending this claim against *Self-doubt, Grief,* and *Explicit prejudice* is to say that:

> *S* only recognises evidence that not-*P as* evidence that not-*P* if they form the belief that not-*P*.

then the conditions on evidence recognition and belief update are trivial. The agent's recognition of their evidence is analysed just in terms of their acquisition of the appropriate state, whilst their acquisition of the appropriate state is analysed just in terms of their recognition of the evidence. But this much is true of implicit bias too. For instance, we can just say that those agents who successfully modified their implicit biases in virtue of reading a single propositional instruction (Gregg *et al.*, 2006) or a strong argument (Briñol *et al.*, 2008) recognised the evidence

that favoured the attitude change. Those who fail to update their biases fail to recognise the evidence. So, if we interpret Gendler's evidence update claim as descriptive, rather than normative, then we lose the distinction between beliefs and implicit biases.

So, if the SD theorist wishes to maintain that there is a substantial distinction between beliefs and implicit biases on the basis that for the former, but not for the latter, recognition of evidence is sufficient for attitude update, then the onus is on them to (1) come up with a non-trivial account of evidence recognition that is independent of the notion of appropriate belief acquisition; (2) explain why the agents in *Self-doubt, Grief,* and *Explicit prejudice* fail to count as recognising evidence; and (3) explain why participants in experiments where implicit biases are modulated in virtue of propositional instruction do not count as updating their attitudes in light of the evidence.

I anticipate that at this stage in the dialectic, we will naturally progress to discussing the agent's *control* of their attitudes, and so we move somewhat beyond the focus of this chapter. I discuss the extent to which we control our beliefs and our explicit biases in the following chapter, where I will pick up again on this debate.

4.3.2. The normative interpretation fails

On the normative interpretation, Gendler's claim is that beliefs are governed by some set of norms which require, but do not guarantee, that they will update in accordance with new evidence, whilst implicit biases are governed by no such norm. I am not convinced, however, that whatever the belief norms turn out to be, they will not *also* apply to implicit attitudes.

To see this, consider arguments that it is just part of the nature of what beliefs are that they aim at truth (for instance, see Williams, 1973; Velleman, 2000). Even when beliefs end up being false, their aim is to represent the world accurately, in accordance with the evidence. On this position, being governed by an evidence-update norm just requires that the mental state in question aims to represent the world accurately. On this 'constitutivist' interpretation of Gendler's claim, the fundamental distinction between beliefs and implicit biases is that only the former constitutively aim at representing the world accurately.

While it may well be true that there is a sense in which beliefs constitutively aim at representing the world accurately, it is not clear why we

should accept this as a substantial distinction claim. Although implicit attitudes might occasionally end up misrepresenting the world (for example, in the case of implicit biases, and Gendler's other examples of alief states), on the whole, implicit attitudes may nevertheless still *aim* at representing the external world accurately, just as beliefs do. Whilst implicit biases have received much philosophical attention, implicit attitudes more generally are implicated in a range of everyday information processing, with some arguing that much daily cognition is implicit, insofar as it occurs and guides behaviour automatically (Bargh & Morsella, 2008). A number of theorists tell an evolutionary story about why we have such automatic processes in the first place, arguing that the evolution of an automatic behavioural guidance system, which represents the world broadly accurately, is, in the context of selective environmental forces, predictable *a priori* (Bargh & Morsella, 2008: 75, see also Dawkins, 1976; Dennett, 1991 & 1995): A creature whose mental states consistently misrepresent their environment, and then figure in behavioural guidance is unlikely to be favoured by selective forces. So, if implicit attitudes evolved with the function of accurately representing the environment, then they too constitutively aim at truth—even if a subset of implicit attitudes (implicit biases) fail to represent truthfully. It is not clear why implicit attitudes wouldn't then be governed by evidence-update norms, just as beliefs are. Not all implicit attitudes *in fact* comply with the norm, because some implicit attitudes end up representing the world inaccurately. However, an attitude's being false does not necessarily imply that it is not governed by a truth norm—constitutivists hold that false *beliefs* are nonetheless still governed by the relevant norm.

An opponent might argue that aiming at truth is a necessarily conscious, necessarily intentional activity. But this need not be the case. Consider Velleman in the following passage:

> A person can also aim cognitions at the truth without necessarily framing intentions about them. Suppose that one part of the person—call it a cognitive system—regulates some of his cognitions in ways designed to ensure that they are true, by forming, revising, and extinguishing them in response to evidence and argument. Regulating these cognitions for truth may be a function for which the system was designed by natural selection, or by education and training, or by a combination of the two. In any case, the

> system carries out this function more or less automatically, without relying on the subject's intentions for initiative or guidance. (Velleman, 2000: 253)

Consider also the three agents from the 'Lights', 'Flat warming' and 'Mushrooms' examples in Chapter 3 who formed beliefs which accurately represented the environment without consciously intending to do so. Presumably, these beliefs were still norm-governed. So, claiming that attitudes may only aim at truth in virtue of conscious and intentional input from the agents in question isn't a commitment of all constitutivists, and will end up ruling out a number of *beliefs* as norm-governed.

So, implicit biases (since they are implicit attitudes) may constitutively aim at truth, and so be governed by evidence-update norms in the same way that beliefs are. Accordingly, the normative interpretation of Gendler's distinction between beliefs and implicit biases fails.

4.4. Implicit Biases As 'Patchy Endorsements', (Levy, 2015)

As I mentioned in Chapter 2, Levy revised the position that he took in his 2014a book, that implicit biases are necessarily associative, in response to Mandelbaum's (forthcoming) argument in a more recent 2015 paper. There, Levy acknowledges the evidence that I summarised in §4.2—that some implicit biases, and implicit social attitudes more generally, have been shown to feature in inferential transitions, which they could only do if they encoded, and were sensitive to, propositional information. He also acknowledges that this refutes arguments that he has made in earlier work which rely on the claim that *all* implicit biases are associatively structured and processed, (2015: 817).

However, Levy (2015) proposes that there is still a distinction between implicit biases and beliefs on the basis of how each are processed, arguing that beliefs are 'inferentially promiscuous' and responsive to evidence:

> Beliefs are *inferentially promiscuous* and beliefs are *responsive to evidence*. Beliefs are inferentially promiscuous inasmuch as the belief that *p* can interact (appropriately) with any other propositional attitude… For instance, my belief that it is raining will interact appropriately with my desire to stay dry, as well as my belief that roads can be dangerous when wet, and any other of my attitudes concerning water and wetness. Whereas inferential promiscuity is a matter of how beliefs cause behavior and update other

> mental states, responsiveness to evidence is a matter of how the belief itself can be expected to update, given appropriate evidence. Inferential promiscuity and responsiveness to evidence are two sides of the same coin: beliefs are inferentially promiscuous, causing the update of other beliefs, because beliefs are responsive to evidence. (Levy, 2015: 805)

Here Levy accepts that the experiments which I summarised in §4.2 show that some implicit attitudes are sensitive to some propositional information. However, he argues that it is significant that many implicit attitudes are *in*sensitive to a lot of other propositional information. For example, he mentions findings which show that implicit attitudes predict job candidate preferences, where subjects do not acknowledge biased preferences in the justification of their choices.[42] He argues that

> …any inference from a proposition like "a white (male) candidate is superior" to "the kinds of qualifications possessed by the white (male) candidate are the ones relevant to the job" is an inference—if indeed it can be called that at all—that ignores too many other representations which we can justifiably attribute to the person. (Levy, 2015: 814)

The claim here is that although implicit attitudes may update in accordance with some propositional information, they fail to update in light of many other propositions. Notably, they fail to update in accordance with the propositional information encoded in the subject's explicit egalitarian attitudes as regards fair hiring practices, for instance. So, whilst implicit biases may be inferentially sensitive, they are not inferentially *promiscuous*, as beliefs are.

Levy proposes that implicit attitudes are a kind of *sui generis* state, that he calls 'patchy endorsements':

> Implicit attitudes are not beliefs. They do not feature often enough and broadly enough in the kinds of normatively respectable inferential transitions that characterize beliefs. Nor, though, are they just associations. They do not activate contents solely associatively: they exhibit some of the kind of

[42] For example, in Chapter 1 we saw that Swedish recruiters with high IAT race bias are significantly less likely to offer a job interview to an applicant with a name that they perceive to belong to a Muslim, as compared compared to applicants with a Swedish name (Rooth, 2007).

> inference aptness that characterize beliefs. They do so in a patchy and fragmented manner, which indicates they have propositional structure. They are patchy endorsements. (2015: 816)

So, whilst Levy acknowledges that at least some implicit attitudes *are* propositional in structure, he maintains that this is insufficient evidence for something like a continuum thesis of implicit attitudes, where implicit attitudes are not fundamentally different to beliefs. He thereby rejects the line of argument that I presented in §4.2.

Levy suggests that there are moral upshots to the notion that implicit attitudes are patchy endorsements (that is, neither purely associative, but also neither fully doxastic) in the following:

> We should hesitate before we blame, or feel shame, or guilt. Equally, though, given that they do not seem to be just associations, there may be room to develop analogues of our existing moral concepts that can apply to agents who harbor them. Right now, neither blame nor excuse (insofar as excuse rests on the claim that they are just associations (Levy 2014a)), seem justified. (2015: 816-7)

In the next subsection I will respond to this conception of implicit biases as patchy endorsements. I will also consider the upshot for moral responsibility, should it be the case that implicit biases are patchy endorsements, when I consider moral responsibility more generally at the end of the chapter.

4.4.2. Inferential promiscuity designates a difference in degree, not kind

I will argue that Levy's distinction between the inferential promiscuity of beliefs and the inferential sensitivity of at least some implicit attitudes is one of degree, not kind, and so supports a continuum thesis, rather than a substantial distinction thesis. I will further show that, in fact, there are some beliefs which are evidence sensitive to a *lower* degree than the most evidence sensitive implicit attitudes. The conclusion will be that Levy's notion of implicit attitudes as 'patchy endorsements' does not uphold a substantial distinction between beliefs and implicit biases.

In the previous subsection, we saw Levy concede that implicit attitudes are structured propositionally. However, he argued that only *beliefs* are inferentially

promiscuous in that "the belief that *p* can interact (appropriately) with any other propositional attitude" (Levy, 2015: 805). Levy intends the distinction between the mere inferential sensitivity of implicit attitudes and the full blown inferential promiscuity of beliefs to designate a substantial distinction in kind. His comments in the abstract, about implicit biases being a *sui generis* mental state, make this clear:

> In this paper I argue that while implicit attitudes have propositional structure, their sensitivity and responsiveness to other mental representations is too patchy and fragmented for them to properly be considered beliefs. Instead, they are a sui generis kind of mental state, a state I dub patchy endorsements. (2015: 800).

Later claims support this too:

> Any state which is inferentially promiscuous and appropriately responsive to evidence is a belief; accordingly, I will follow Brownstein and Madva (2012) in taking this *kind* of responsiveness to be the mark of a *bona fide* belief. (Levy, 2015: 805)

Levy (2015) doesn't explicitly say what constitutes the different kinds of ways to update in light of evidence, or the different kinds of ways to be inferentially sensitive, if indeed he holds that such differences exist. Instead, it seems that inferential promiscuity is just *frequent* inferentially sensitivity. But this would make the distinction one of degree, not kind.

Later on in the paper, the argument seems to be that the relevant difference in kind will be constituted by a large enough difference in degree. This is evident in the following passage (where I have underlined words and phrases which I think indicate a difference in degree):

> Both sides would surely agree, however, that excessive evidence insensitivity and encapsulation blocks the ascription of a correlative belief to an agent. … Though it may be ineliminably vague just how much responsiveness to evidence is required for a representational state to count as a belief, sufficient departure from the kind of sensitivity to evidence and

> aptness for normatively respectable inference we associate with a *bona fide* belief will settle the question. (Levy, 2015: 806, underline emphasis mine)[43]

So, according to the above passage, the idea seems to be that, at some point, a distinction in degree becomes a distinction in kind, even if the precise extent to which an attitude must be responsive to evidence in order to count as a belief, rather than implicit, is "ineliminably vague".

I think that this commits the patchy endorsement theorist to the following claims:

PE1: Each and every state in the set of beliefs, b_1, b_2, $b_3 \ldots b_n$, is responsive to evidence more frequently than each and every state in the set of implicit attitudes ia_1, ia_2, $ia_3 \ldots ia_n$.

PE2: There is a sufficient gap between the least evidence-responsive belief, b_L, and the most evidence-responsive implicit attitude, ia_M such that b_L is a different *kind* of state to ia_M.

Patchy endorsement theorists are committed to PE1 insofar as they hold that both beliefs and implicit attitudes are propositional, and so some particular belief b_{23} does not respond to some piece of evidence e_1 in a different *kind* of way to the way in which some particular implicit attitude ia_{67} responds to evidence e_1. Instead, the distinction lies in the *frequency* with which b_{23} responds to evidence: for example b_{23} responds to evidence e_1, e_2, e_3, e_4, and e_5, whilst ia_{67}: only responds to e_1 and e_2. In other words, whilst ia_{67} may be inferentially sensitive, b_{23} is inferentially promiscuous. PE1 is the minimum commitment for the patchy endorsement theorist to be able to say that all beliefs are inferentially promiscuous whilst all implicit biases are only inferentially sensitive—whatever these properties turn out to be—in that it orders implicit attitudes and beliefs on a continuum of inferential sensitivity, but does not determine the distance between the most evidence-responsive implicit attitude, and the least evidence-responsive belief. PE2 is the stronger commitment which delivers the desired difference in

---

[43] Although we've switched from 'inferential promiscuity' to 'responsiveness to evidence', recall that Levy thinks that "Inferential promiscuity and responsiveness to evidence are two sides of the same coin: beliefs are inferentially promiscuous, causing the update of other beliefs, because beliefs are responsive to evidence" (2015: 805).

kind, even though, as we saw Levy admit above, this distinction will fall at an arbitrary point on the continuum of inferential sensitivity to inferential promiscuity.

PE2 requires, as a minimum, that PE1 is true. So, if PE1 is false, this entails that PE2 is false. Given this, we can test the patchy endorsement claim that implicit attitudes are a *sui generis* class, distinct from beliefs: If we can find even one belief that is less frequently responsive to evidence than the most frequently evidence-responsive implicit attitude, then PE1 fails. If PE1 fails, then PE2 fails, and consequently, the patchy endorsements argument, that there is a substantial distinction between implicit attitudes and beliefs on the basis of the frequency of responsiveness to evidence, fails. In the following, I argue that PE1 fails, and PE2 fails also.

Firstly, it is unclear what motivates PE1. Admittedly, there is evidence that some implicit attitudes, at least sometimes, do not update in response to evidence, or that they update, but not in the right way (Levy, 2015). But there are also cases where some *beliefs*, at least sometimes, do not update in response to evidence, or that they update, but not in the right way (such as in the examples of *Self-doubt, Grief,* and *Explicit prejudice,* from §4.2.2). Recall that nothing in the nature of implicit attitudes suggests that they will consistently misrepresent the environment, or else creatures with cognitive systems which utilise implicit attitudes are unlikely to be favoured by selective forces.

To the further detriment of PE1, I think that there is a whole class of beliefs which are at least as frequently unresponsive to evidence as the most responsive of implicit attitudes: that is the class of *explicit* prejudices. In fact, explicit prejudices which have been held for a long period of time persist precisely because they are frequently unresponsive to evidence. To see this, consider Liz, who believes that "black people are unintelligent." Liz is introspectively aware of the content of her belief: She feels assent to it, taking it to describe a truth when she recalls the content. She is disposed to assert it—not in all circumstances, for example, not in front of her boss—but she's happy to bring the content of her belief up with friends, when in the pub, and in conversation with her boyfriend. It guides her actions in appropriate ways, leading her to utter racist slurs, and to assert "Too thick to learn English" when a black footballer makes a grammatical mistake in a post-match interview. However, when Liz encounters evidence which counts against her belief, she fails to update her belief

appropriately in light of this evidence. She fails to update her belief when she reads about a black oncologist developing a new drug which significantly enhances the effectiveness of leukemia treatment. She fails to update her belief when a black colleague fixes a bug in the software her team is designing. She fails to update her belief when a black friend of a friend beats her at chess three times in a row. Liz has plenty of evidence, and plenty of opportunity to update her prejudiced belief about black people. And yet, she fails to: her belief not only fails to be inferentially promiscuous in light of this evidence, but it fails to be even moderately inferentially sensitive.

Resistance to evidence seems to be an important part of how we characterise at least some prejudice. Consider Arpaly's example of Soloman. Soloman comes from an isolated village in which he is not exposed to any women who are also abstract thinkers, and so forms the belief that women are incapable of abstract thought. But then:

> Imagine…that Solomon gains a scholarship and finds himself a student in an excellent academic institution, where he proceeds to study his favorite abstract topic. In college, Solomon sits shoulder to shoulder with brilliant female students and is taught by brilliant female professors. At the end of his first year as a college student, if Solomon were rational, he would have changed his mind about the aptitude of women for abstract thinking. If at the school year's end Solomon still believed that all women are bad abstract thinkers, his belief would now be not only false but also irrational. He would no longer be simply mistaken, but *prejudiced*. (Arpaly, 2003: 104)

Resistance to a year's worth of evidence that women clearly are capable of abstract thought, and yet *failing* to update his belief in light of all of this evidence is the very fact that renders Solomon's belief a prejudice. In fact, it would seem that we can only explain why explicit prejudices persist by characterising them as regularly inferentially insensitive.

So, the least evidence-responsive belief has set the bar of responsiveness to evidence—a bar which, for the patchy endorsements theorist, must not be exceeded by *any* implicit attitude—rather low indeed. It seems entirely possible that a person could remain prejudiced for a lifetime, in which case, this would be an example of a belief which is almost *entirely* insensitive to evidence. It seems highly unlikely that the most evidence sensitive implicit attitude is still less

evidence sensitive than this. We considered in §4.3.2 that a creature whose implicit attitudes consistently misrepresent their environment is unlikely to be favoured by selective forces (Bargh, 2008: 75). A creature whose implicit attitudes are *all* less evidence sensitive than Liz's explicit prejudice seems very unlikely to be favoured in the face of selective evolutionary forces. Further, we've already seen a number of empirical findings which show that at least some implicit attitudes *do* update appropriately in response to evidence.

So, PE1, the claim that every belief is responsive to evidence more frequently than every state in the set of implicit attitudes, fails. If PE1 fails, then PE2, the claim that there is a sufficient gap between the least evidence-responsive belief, and the most evidence-responsive implicit attitude to constitute a difference in kind between beliefs and implicit attitudes, also fails. At best, the patchy endorsements theory is a theory about the difference in the *degree* of evidence sensitivity of some implicit attitudes as compared with that of some beliefs. If so, that would make it a continuum thesis. It is no basis for a substantial distinction between all beliefs and all implicit attitudes.

SUMMARY

In the forgoing, I argued that there is no significant distinction between (i) the structure of implicit biases and the way in which they are processed; and (ii) the structure of beliefs and the way in which they are processed. I argued that at least some of our implicit biases are propositional in structure, and feature in evidence-sensitive inferential transitions in the same way that many beliefs do. I demonstrated that SD claims on the basis of structure and processing generate two testable hypotheses. I showed that HYPOTHESIS 1 (that implicit biases are necessarily associative) is false, by summarising findings from both de Houwer (2014) and Mandelbaum (forthcoming) in which some implicit biases are shown to be sensitive to propositional information. I then argued that this finding demotivates the expectation that implicit biases are generally associative. Following this, I argued that HYPOTHESIS 2 (the descriptive interpretation of the claim that beliefs change in response to changes in evidence) is false, by providing a number of examples of beliefs which fail to update in accordance with new evidence. I then argued that even if the claim from HYPOTHESIS 2 is interpreted normatively, it will fail to distinguish between implicit biases and beliefs, because implicit attitudes may be governed by the same kind of epistemic

norms as beliefs. Finally, I considered Levy's (2015) reinterpretation of HYPOTHESIS 2, on which implicit biases are propositionally structured 'patchy endorsements', but which, unlike beliefs, fail to be inferentially promiscuous. I argued that this account fails to reinstate the substantial distinction between implicit attitudes and beliefs on the basis of a difference in processing, demonstrating that there are at least some beliefs that are evidence sensitive to a *lower* degree than the most evidence sensitive implicit attitudes.

The argument made in this chapter undercuts SDR arguments which proceed on the assumption that there is substantial distinction between the structure of implicit biases, and the way in which they are processed, and the structure of beliefs, and the way in which they are processed. In light of this, consider the following claims of SD theorist Gendler (2008a, 2008b) summarised in TG1, and SDR theorist Levy (in his 2014a book) summarised in NL5 and NL8.

TG1: Implicit biases are aliefs: sui generis tripartite mental states with a representational component, an affective component, and a behavioural component, which are 'associatively linked'.

NL5: Implicit biases are associative, not propositional, in structure.

NL8: The fact that I associate X and Y, nonconsciously, is no basis for holding me morally responsible.

Whether or not NL8 is correct, since NL5 and TG1 were shown, empirically, to be false, (a result that Levy himself later acknowledges in his patchy endorsements paper, 2015: 817), NL8 tells us nothing about implicit biases. In fact, NL8 is consistent with moral responsibility for at least some implicit biases and the actions that they influence. Consistency of course does not show that we *are* morally responsible for harbouring any implicit biases, or for any implicitly biased actions—just that this isn't ruled out on the basis of the attitudes' structure.

In light of the arguments made in his 2015 paper on patchy endorsements, Levy makes a slightly different suggestion as regards moral responsibility for implicit attitudes and the actions that they influence at the end of this paper, to that which he argued for in the 2014a book. Even though the main argument in the 2015 paper is that implicit biases are a *sui generis* class, distinct from beliefs—an SD argument, Levy does not follow this up with an SDR argument to the effect that we are therefore not in *any* way morally responsible for harbouring implicit biases, or for actions influenced by them. He appears to concede that such

an SDR argument is not going to follow straightforwardly from the notion that implicit biases are patchy endorsements, acknowledging that "[w]e should hesitate before we blame, or feel shame, or guilt for habouring implicit biases, or for their influence on action" (Levy, 2015: 816-7. Note that this is a much weaker claim than the SDR claim of the 2014a book).

It is not an argument that Levy (2015) makes in any detail, but let me anyway anticipate a possible SDR argument on the basis that implicit biases are not inferentially promiscuous, and offer a continuum response. This would be an argument along the lines that there is some level of inferential promiscuity that it is necessary for an attitude to have in order for us to be morally responsible for that attitude and the actions that it guides. The SDR theorist would hold that beliefs exceed this level of inferential promiscuity, whilst implicit biases fall below it. I think that this argument will fail. Firstly, because there are at least some beliefs that are evidence sensitive to a *lower* degree than the most evidence sensitive implicit attitudes, we cannot draw a line on the implicit bias-belief evidence sensitivity continuum, such that we are morally responsible for all beliefs, and the actions that they guide, and for no implicit biases, and the actions that they influence. We will either end up excluding at least some beliefs and belief guided actions as possible targets of moral condemnation, or *including* at least some implicit biases and implicitly biased actions.

Secondly, the degree of evidence sensitivity of an attitude is not obviously related to moral responsibility, at least in the case of agents with a recalcitrant explicit prejudice, such as Liz. Recall that Liz's explicit prejudice not only failed to be inferentially promiscuous in light of this evidence, but failed to be even moderately inferentially sensitive. But nonetheless, Liz seems to be a target for moral condemnation *precisely because* she has failed so frequently to update her racial prejudices in light of counter-evidence. The very trait which, according to Levy (2014a) exempted implicit biases, and their manifestation in action, from moral condemnation seems to be the very trait that is morally repugnant about Liz. We might think that the moral condemnation that we feel towards Liz is not wholly a function of the fact that she has failed to update her prejudiced attitudes, but it is also a function of her awareness of the situation. That seems reasonable. But it is not a way to save the particular SDR theory at hand which says that inferential promiscuity is a *necessary* condition for moral responsibility.

So, if it turns out that we do lack moral responsibility for all of our implicit biases, and their influence on our actions, then it will *not* be because of their structure, and the way in which they are processed, but because of some other distinguishing feature. In §4.3.2, I reached a stage in the dialectic where the natural progression was to discuss agential control. SD(R) claims on the basis of control will be my focus in the next chapter.

CHAPTER 5: RESPONDING TO SUBSTANTIAL DISTINCTION CLAIMS ON THE BASIS OF CONTROL

In the previous two chapters, we saw that a fundamental distinction in kind between implicit biases and beliefs (and their associated actions), could not be upheld on the basis of the kind of awareness that we have of each (Chapter 3) or on the basis of the structure and processing of each (Chapter 4). This chapter examines the final set of arguments for the substantial distinction account of implicit bias: arguments on the basis of control. I argue that there are a number of strategies that we can utilise to control implicit biases and implicitly biased actions—strategies which, as I will demonstrate, are also necessary for controlling belief acquisition, and at least some of our everyday agential actions. The conclusion will be that there is no substantial distinction between implicit biases and agential attitudes on the basis of the kind of control that we exert over each and their associated actions.

I start by providing an overview of some of the notions of control that will be relevant to the discussion (§5.1). I briefly introduce the notion of control in the psychological literature (§5.1.1), which will in part inform how we are to understand SD claims on the basis of psychological findings. I then introduce some of the main accounts of agential control in the philosophical literature: (i) voluntary control; (ii) reasons responsiveness; and (iii) deep self accounts. I also define a number of important distinctions: (a) direct *vs.* indirect control; (b) initiation *vs.* intervention control; and (c) deliberative *vs.* non-deliberative control (§5.1.2) which will be relevant to the demonstration that we do exert several kinds of control over implicitly biased actions. Following this, I outline the substantial distinction claims on the basis of control that were presented in Chapter 2 (§5.1.3). Accordingly, the relevant SD claims require that there is no kind of control that we have over at least some of our implicit biases, and their guidance of action, that it is necessary for us to utilise in order to control our beliefs, and their guidance of action.

In §5.2, I consider SD claims as they relate to the acquisition and maintenance of implicitly biased *attitudes vs*. that of beliefs. §5.2.1 explores a strategy apparently open to continuum theorists, who may be tempted to utilise Bernard Williams' (1973) claims that we do not exert control over the acquisition

of beliefs to argue against any substantial distinction between them and implicit biases. I argue that this strategy is unsuccessful, and shows only that we lack *direct* voluntary control of beliefs. An SD argument may be advanced instead on the basis that we have *indirect* voluntary control of belief but lack such control of implicit bias. I show, however, that the SD argument on the basis of indirect voluntary control fails, by presenting evidence to the effect that we also have indirect control over the acquisition and maintenance of at least some of our implicit biases (§5.2.2).

I then consider another response to Williams (1973): that of Pamela Hieronymi in her (2008) notion of 'answerability', according to which we to exert a kind of direct control over our beliefs because they embody our take on the world (§5.2.3). One might think that this account reinstates the substantial distinction. I demonstrate, however, that it does not. SD theorists may try to advance an argument on the basis of answerability, with the effect of showing that we lack direct control of the acquisition of our implicit biases. But they will be unable to avoid *also* showing that we lack direct control of the acquisition of at least some of our beliefs. So, the argument that there is a substantial distinction between beliefs and implicit bias, in terms of the control that we exert over their acquisition, fails.

There are other ways in which one may attempt to advance the substantial distinction theory on the basis of control: for instance, in terms of control over *actions* guided by beliefs *vs* those guided by implicit biases. This is the focus of §5.3. I show that we in fact have indirect, intervention control of many of our implicitly biased actions (§5.3.1), and, further, that indirect, intervention control is the *only* kind of control we may exert over many uncontroversially agential actions. But indirect control is not the only kind of control that we have over implicitly biased action: agents may exercise two kinds of *direct*, intervention control over the manifestation of implicit bias in action, as I will demonstrate in §5.3.2. In particular, agents have (i) a form of deliberative, direct, intervention control; and (ii) a form of non-deliberative, direct, intervention control. I demonstrate that we rely on both (i) and (ii) in everyday agential action. As regards (i) I argue that we can—and are often *expected to*—directly and deliberately intervene on the manifestation of a number of more familiar preferences in daily life to render our agential outputs more effective, expressive or fair and, therefore, this method of control is indispensable to the guidance of

some of our everyday agential actions. As regards (ii) I outline some research which suggests the neural basis for a more general faculty for non-deliberative, direct, intervention control that may well operate to enable us to control our actions in accordance with our motivations in a number of everyday activities.

I thus demonstrate, in §5.3, that there are three different control strategies which are effective over our implicitly biased actions, which are also *necessary* for controlling at least some of our everyday agential actions. This result is inconsistent with the SD theorist's claim that there is a kind of control that we have over all of our belief-guided actions which we do not have over any of our implicitly biased actions. As such, SD arguments as regards control of actions, fail.

We will then be in a position to consider the SD claim that the (apparent) fundamental distinction in control over implicit biases and agential attitudes (and the actions that they guide), rules out moral responsibility for implicit biases and implicitly biased actions. At this point, I will also consider how awareness interacts with the control that we exert over implicitly biased attitudes and actions, should the (apparent) lack of awareness *and* the (apparent) lack of control be jointly sufficient support for the SDR claim that we lack moral responsibility for implicit biases and the actions that they guide. I argue that, if it turns out that we do lack moral responsibility for our implicit biases, and their influence on our actions, then it will not be because we lack awareness of, or control over them. This result will pave the way for a more positive statement of the continuum thesis in Chapter 6.

5.1. Control: Psychological & Philosophical Notions

Philosophers and psychologists do not necessarily always understand 'control' to mean the same thing, and so arguments sometimes may be at cross-purposes. In order to understand exactly what claims are being made by substantial distinction theorists who utilise psychological results, we need to identify which account of control they rely on. In the following subsections, I give an overview of some psychological notions of control and their relevance to research on implicit bias (§5.1.1); some philosophical accounts of control, and distinctions which will be relevant to the discussion to come (§5.1.2); and a recap of the substantial distinction arguments on the basis of control, that I will address in this chapter (§5.1.3).

5.1.1. *Psychological notions of control*

Recall from §1.3 of Chapter 1, the notion that there exist 'automatic' processes which contrast with 'controlled' processes was proposed in the psychology literature by Schneider and Shiffrin (1977), who contributed to research into attention and mental processing from which the paradigms to test for implicit bias eventually evolved. Schneider and Shiffrin define an automatic process as that which is activated "without the necessity for active control or attention by the subject" (1977: 2). Later psychologists contrast automatic or uncontrolled processes with those processes which involve the subject's intending to fulfil a task or to achieve a goal (Moors *et al.*, 2010: 20). In particular, according to Moors *et al*., it is characteristic of an automatic process that it can result in particular effects when the subject in question did not have the goal of achieving such an effect (2010: 20).

A number of psychologists make use of the distinction originally proposed by Schneider and Shiffrin (1977) to analyse implicit biases, suggesting that implicit biases mediate actions via automatic, as opposed to controlled, processes: Dasgupta and Greenwald, in discussion of their 2001 experiment on what they term 'automatic prejudice', write that 'activation of automatic beliefs has been described as an inescapable habit that occurs despite attempts to bypass or ignore it,' (2001: 800. Dasgupta & Greenwald attribute this thought, in particular, to Bargh, 1999, and Devine, 1989). In Chapter 1, I already mentioned evidence that participants are unable to respond more quickly on stereotype incongruent trials just because they are instructed to do so, (Banse, Seise, & Zerbes, 2001); as well as evidence that participants who are asked to form a goal to not stereotype are not able to reduce the extent to which their responses are stereotyped on an IAT (Lowery *et al*. (2001). Another argument for the notion that implicitly biased actions are not controlled is the idea that we take it that people act consistently with the values which they take themselves to have, or at least, values which they profess to have on self-report measures (Nosek *et al.*, 2007).

For others, controlled processes are those which involve the rule-based processing of propositions, whilst uncontrolled processes are those which involve the processing of associations, (Gawronski & Bodenhausen, 2014). I presented evidence and arguments against this sort of dual process model—insofar as it

applies to implicit bias—in the last chapter, and so will not discuss accounts of control on the basis of rule-based processing further here.

Let us now turn to philosophical accounts of control, where we see a proliferation of notions, to which the psychological notions described above map neither obviously nor neatly.

5.1.2. *Philosophical accounts of control*

The philosophical literature on agential control—what it is for an agent to be in control of an attitude or an action—is vast. But we need to have at least a basic understanding of this literature to make sense of the relevant SD claims: to assess whether or not they are supported by the psychological evidence, and to investigate whether they really do deliver a substantial distinction. Some philosophers have suggested that an agent is in control of some act if they are able to act voluntarily. Others have suggested that an agent is in control when they are able to respond to reasons. Yet others still have suggested that agents are in control when their action expresses their 'deep self' in a sense that I will define shortly. That said, I do not have space to offer anything more than a brief outline of these philosophical accounts of control. There are further distinctions which cut across these accounts just mentioned, which will be relevant to the discussion in this chapter. They are: direct *vs.* indirect control; initiation *vs.* intervention control; and deliberative *vs.* non-deliberative control. I will give a brief outline of each of these distinctions in the following.

So, we turn to the first philosophical account of agential control.

Voluntary control

A number of philosophers maintain that agential control is to be understood in terms of the operation of the agent's will. For instance, an agent has voluntary control over some ϕ-ing when when her ϕ-ing is the outcome of her will. Famous proponents include Descartes (1694/1984) and Mill (1843/2002). There is a large literature on what the will amounts to (for discussion, see O'Connor, 2005; Hyman, 2015: Chapter 1). For many voluntary control theorists, "conscious choosing" is the paradigm output of the will (Hyman, 2015: 2). For others, to will is to exercise a capacity wherein it is possible for an agent *not* to do as they do (Kenny: 1963: 237). Yet others theorists who espouse voluntary control theories identify willing with fulfilling one's intentions (Mele, 1992; Mele and Moser,

1994). The notion of voluntary control will be relevant to the discussion to come. However, I will not adopt any particular account. Instead, my purpose will be to show that whatever kind of voluntary control we are supposed to have over our agential attitudes and actions, we also have it over at least some of our implicit attitudes and actions, and so substantial distinction arguments on the basis of voluntary control do not work.

Reasons responsiveness

Another tradition is to characterise some ϕ-ing as controlled and agential only if in so ϕ-ing the agent was able to respond to (practical) reasons for ϕ-ing. Notable proponents in this tradition are Wolf (1990) and Fischer and Ravizza (1998). These accounts hold that in at least some circumstances, as well as acting for reasons they see there to be for acting, agents must also be able to recognise and to respond to reasons for *not* acting as they do in at least some circumstances.[44] Reasons may be considerations that justify acting from the agent's perspective, or they may be features which explain an agent's actions. Justificatory and explanatory reasons do not necessarily always converge (Dancy, 2000). Some philosophers argue that reasons are facts about the agent's situation, whilst others argue that reasons are internal mental states (see Alvarez, 2009), but my discussion in this chapter will be neutral on this issue. As was the case with voluntary control, my purpose is not to adopt any particular account of reasons-responsiveness, but to show that, however SD theorists understand a capacity to respond to reasons, it will not support a substantial distinction between beliefs and implicit biases (and actions guided by each).

Deep self

Others think that agents control just those ϕ-ings which manifest their fundamental evaluations the world—their 'deep self' (see Watson 1996; Smith 2005, 2008, 2012, manuscript; Sher, 2006; Hieronymi, 2008; Sripada, forthcoming). In particular, deep self theories can provide an account of why agents are implicated in cases of omission (which are often neither voluntary, nor

[44] Fischer and Ravizza (1998) argue that agents need to be able to do this for all controlled instances of action, whilst Wolf (1990) maintains that agents only need to be able to recognise and to respond to reasons for *not* acting in circumstances where they do something 'unreasonable': she relies on an objective sense of reason to delineate what is reasonable and unreasonable.

done for reasons), maintaining that omissions nevertheless reflect the things that the agent really cares about—or fails to care about (for example see Smith, 2005; Sher, 2006). Not all proponents of deep-self accounts consider them to be theories of control *per se*, some understanding the term 'control' to implicate a voluntary faculty. However, for deep-self theorists, agents play an active role in ϕ-ings which manifest their evaluative stance, whether or not they also ϕ voluntarily, and so on a neutral notion of what control amounts to, deep-self accounts are worthy of inclusion. As above, I do not intend to adopt any particular deep self account in what follows, but to demonstrate that however substantial distinction theorists cast the deep-self, it will not enable them to maintain the fundamental distinction that they require.

Variation and interdependence across accounts

There is great variation within each of these broad traditions, but also, it is worth noting that some accounts within one tradition appeal to notions from another to analyse agential control. For instance, some have analysed voluntariness *as* responsiveness to reasons (such as, for instance, Bennett, 1990: 90). Further, Hieronymi (2008) presents a deep self account of doxastic control, arguing that our beliefs manifest our evaluative stance, but her analysis of what this amounts to appeals to the notion that it is appropriate to ask the agent to give *reasons* for her beliefs. I will consider Hieronymi's (2008) account in more detail in §5.2.2.

There are a number of other distinctions which cut across all three accounts presented above, and each other, that are relevant to the argument in this chapter. I think that previous philosophers in this literature have conflated some distinctions, that I will now tease apart. Later, in §5.3, we will see how these particular distinctions enable the continuum theorist to outline three distinct kinds of control that we have over our implicitly biased actions, which are also the *only* kinds of control that we have over at least some of our belief-guided actions, which is bad news for the SD theorist.

Direct vs indirect control

Many have distinguished between 'direct' and 'indirect' notions of control (Williams, 1973; Bennett, 1990; Feldman, 2001; Strawson, 2003; Hieronymi, 2008). Accordingly, an agent has direct control of some ϕ-ing if it is within their power to ϕ, without any intermediary steps. This is to say that the agent's bringing

about her ϕ*-ing* is *itself* the act of agential control. It would seem that many bodily movements are under this type of control. For example, an agent has direct control over raising her arm, if it is within her power to raise her arm without any intermediary steps. Pollard captures this idea in the following (although, shortly, I will argue that there are at least three different distinctions in kinds of control to draw out of Pollard's (2003) account).

> When we deliberate we exert a kind of *direct* control over what we do: we think about what to do, and then do it. (Pollard, 2003: 415)

In the case of direct control, the action that we seek to bring about, (and will to do, or intend to do, or see there to be reason to do, and so on—depending on your chosen account of control) is an action that it is within our power to perform without having to take preparatory steps.

In the case of indirect control, the act which the agent themselves performs is a *distinct* event from that which she seeks to bring about. Suppose that the agent wants to bring about ϕ. She can bring about ϕ in an act of indirect control if she can bring about ψ (through an exercise of *direct* control, as above), and from the occurrence of ψ, the occurrence of ϕ follows. The cases of indirect control that I am interested in are those where ϕ is a change in the agent's attitudes or actions.[45] For instance, it may be argued that it is not within my direct control to make myself feel happy, just like that. However, I do have direct control over seeking out and looking at a photo of an enjoyable holiday, or recalling a memory of time spent with a loved one, and as a result of presenting myself with these happy memories, I feel happy. The recall of a joyful occasion is appropriately linked up in my psychological economy to the occurrent experience of happiness such that I may experience the latter as a result of recalling the former. So, even if I cannot simply bring it about that I am happy in the same way that I can raise my arm, I can look at a photo, or recall a joyful experience, and as a result of so doing, indirectly bring it about that I am happy.

The relevance of this distinction to the case of implicit bias is that whilst one might think that we lack direct control over our implicitly biased actions (for

---

[45] One might have indirect control over all sorts of external objects, in virtue of being able to, through an exercise of direct control, bring about a change in them. But let us set these sorts of cases aside, and focus on cases where ϕ is a change in the agent's attitudes or actions.

instance) it may be that we are able to control them in virtue of doing something else which itself has the effect of preventing the manifestation of a biased action. For example, anonymising C.V.s prevents the manifestation of bias against applicants that might arise as the result of viewing gendered or racialised names. Here, the act of anonymising is distinct from the act in which we seek to eliminate bias (that is, the evaluation of the applications) but elimination of bias in evaluation is the result of the earlier anonymising. It is also worth flagging up that I think that there are at least two kinds of direct control that we have over implicitly biased actions, as I will discuss more fully in §5.3.2.

Initiation vs intervention control

The quotation from Pollard in the previous section comes from a 2003 essay on automatic actions, in which he outlines a slightly different notion of indirect control to that which I outline above. He maintains that we have such control over at least some of our automatic behaviours. He calls this notion 'intervention' control. He suggests

> direct control is absent when our behaviour is automatic…we have the capacity to intervene on such behaviours. This is particularly the case for those automatic behaviours which we have learned. Since there was a time when we didn't do such things, it will normally still be possible for us still to refrain from doing them… We intervene by doing something else, or nothing at all, either during the behaviour, or by anticipating before we begin it. (Pollard, 2003: 415)

For instance, I may notice myself performing one of my automatic habitual actions—nail-biting—and intervene on my behaviour by doing something else, namely by ceasing to bite my nails. Snow (2006) suggests that the appropriate complement of Pollard's intervention control is to be termed 'initiation' control:

> As I understand Pollard, direct control could also be called "initiation control," since it is the kind of control we exert when we initiate an action or action sequence. (2006: 549)

Accordingly, I do not generally *initiate* sequences of nail-biting behaviour, even if I can intervene on them. I think that this notion of initiation control is distinct

from the notion of direct control that I outlined above. In fact, I think that, precisely because it is to do with sequences of behaviour that have become automatised *vs.* non-automatic behaviour, the Pollard-Snow initiation/intervention distinction is independent of the direct/indirect distinction as I outlined it above. To see this, recall my example of indirect control in the previous section, where an agent thinks of a happy memory to engender an occurrent feeling of happiness in herself. She is not intervening on a process already in motion, but initiating a new one. However, her control of her emotion is not direct, because she has to perform an act which is itself distinct from feeling happy in order to bring about a feeling of happiness. So, these distinctions are independent and combine to give us four possible types of control:

> (1) I have direct, initiation control over the raising of my hand;
> (2) I have direct, intervention control over the ceasing of my nail biting;
> (3) I have indirect, initiation control over bringing on an occurrence of joyousness, by looking at photo of an enjoyable time; and
> (4) I have indirect, intervention control when I get the (almost uncontrollable) giggles in the library, and I call to mind a sad event in my life, which has the effect of curtailing my giggling.

Case (4) is somewhat like case (3), but my intention is to intervene on some emotion-driven behaviour which is already in motion. In short, we can employ direct control to either (1) initiate or (2) intervene on a sequence of behaviour, and we can employ indirect control to bring about some occurrence which itself either (3) initiates or (4) intervenes on a sequence of behaviour.

The initiation/intervention distinction has been recently employed in the implicit bias literature by Holroyd and Kelly (2016). They develop an account on which agents have a form of indirect, intervention control that is effective over a great many implicitly biased actions. I will present Holroyd and Kelly's account of indirect, intervention control in more detail in §5.3.1 (although, in §5.3.2, I suggest that one of their examples is better characterised as a case of *direct*, intervention control).

Deliberative vs non-deliberative control

The examples in (1)-(4) above are all examples where the agent has deliberated about what to do, and has then done it. By deliberation, I mean something akin to Pollard's 'we think about what to do' (2003: 415) prior to acting, where the action performed is that which is specified in the outcome of our thinking about what to do.[46] Such deliberation does not necessarily need to be prolonged or carefully thought through for subsequent actions to qualify as deliberatively controlled. As long as they were thought about at all, then they may be deliberatively controlled.

Contra Pollard, however, I think that at least sometimes, agents may employ at least some of the kinds of control detailed in the previous two subsections *without* prior deliberation, as I will argue in §5.3.2. Pollard maintains that intervention control is typically deliberative. In particular,

> Of course when we intervene, this will typically require thought, and the behaviour will on that account cease to be habitual. (2003: 41)

I do not think there are many (or perhaps any) instances of non-deliberative, *indirect*, intervention control. But, contra Pollard, I do think that we sometimes have deliberative, *direct*, intervention control, of both agential actions and implicitly biased actions, as I will argue in §5.3.2. I also think that we sometimes have *non-deliberative*, *direct*, intervention control, of both agential actions and implicitly biased actions, as I will argue in that same section.

To recap these three distinctions, then: The direct/indirect distinction refers to whether the ϕ-ing which the agent performs is the occurrence which she seeks to bring about or whether it leads to the occurrence that she seeks to bring about. The initiation/intervention distinction refers to whether the agent initiates a new sequence of ϕ-ing, or whether she intervenes on some ϕ-ing already underway. The deliberative/non-deliberative distinction refers to whether the ϕ-ing is the outcome of deliberation, or not.

---

[46] Holroyd and Kelly make a similar distinction, differentiating 'taking' from 'exercising' control, where the former is the outcome of a deliberative process, but the latter is not, (2016: 121). For consistency, I will stick to the terms 'deliberative' and 'non-deliberative'.

5.1.3. *Substantial distinction arguments on the basis of control*

In Chapter 2, I summarised the main substantial distinction claims which utilise the notion of control. We saw claims about our apparent lack of control over the *acquisition* of implicitly biased attitudes:

K&R5: Implicit biases are acquired rapidly, automatically, and uncontrollably

JS2: We are not able to exert control over the acquisition of our implicit biases, in virtue of the fact that they result solely from our living in a bigoted culture.

JS3: Inferential awareness that we are likely to be implicitly biased is not sufficient for control over implicit biases (interpreted as implicit biases *qua* attitudes)

We also saw claims about our apparent lack of control over the effects of implicitly biased attitudes on *actions*:

K&R4: Implicit biases influence action automatically.

TG2: Activation of the representational content of an implicit bias renders it more likely that an implicitly biased behavioural routine will actually be performed.

JS3: Inferential awareness that we are likely to be implicitly biased is not sufficient for control over implicit biases (interpreted as implicitly biased actions)

As noted in §5.1.2, there are multiple possible notions of control according to which we may interpret the above claims. Clearly, I will not have space to systematically consider whether each of the above claims delivers a substantial distinction on the basis of each and every possible kind of control. However, I aim to examine most possibilities. Further, if it is the case that there is a kind of control that we can exert over at least some implicit biases/implicitly biased actions, which is also the *only* kind of control that we exert over at least some agential attitudes/agential actions, then we have a result that is inconsistent with the SD theory. I will show that this result in fact obtains.

I will distinguish arguments on the basis of the control that we exert over the acquisition of implicit biases, compared with acquisition of agential attitudes

(§5.2), from arguments on the basis of the control that we exert over implicitly biased actions, compared with agential actions (§5.3). In §5.2, I will present SD arguments (and continuum responses) on the basis of indirect voluntary control (in §5.2.2) and direct deep-self control (in §5.2.3). In §5.3, I will present SD arguments (and continuum responses) on the basis of indirect, intervention control (in §5.3.1), and two possible accounts of direct, intervention control: (i) deliberative, direct, intervention control; and (ii) non-deliberative, direct, intervention control (in §5.3.2).

5.2. Control of Implicitly Biased Attitudes

In the last chapter, I looked at a number of claims from SD theorists that beliefs update in light of propositional information, whilst implicit biases are unable to, because they do not have the appropriate structure—that is, they are not propositional. I refuted these arguments with evidence that at least some implicit biases are propositional, and do update in light of new propositional information. I also argued that if at least some implicit biases are propositional in structure, then we lose the motivation to expect that other implicit biases will be associative. I then demonstrated that at least some beliefs fail to update in light of new propositional information, even when the agent recognises the new propositional information as evidence relevant to their belief. I argued that this shows that there is no substantial distinction between implicit biases and beliefs on the basis of their structure and processing. I acknowledged at the end of §4.2 that some SD theorists may reply to these arguments by suggesting that, even though implicit biases might have the right structure to be updated in light of propositional information, we have a kind of control over the acquisition of our beliefs that we do not have over our implicit biases. I now turn to discussion of this claim.

SD theorists who argue that there is a substantial distinction between implicit bias acquisition and belief acquisition need it to be the case that there is a kind of control that we have over the acquisition and maintenance of our beliefs (call this 'doxastic control') which we do not have over the acquisition and maintenance of our implicit biases. In what follows, I first consider an argument from Williams (1973) that we lack voluntary doxastic control, and acknowledge that this may appear to represent a victory for the continuum theorist (§5.2.1). This victory is premature, however, because Williams (1973) only shows that we lack *direct* voluntary doxastic control, and his account is quite compatible with

our having *indirect* voluntary doxastic control, which may reinstate a substantial distinction between implicit biases and beliefs (§5.2.2). However, I respond to the threat of a substantial distinction here, by showing that we also have indirect control over the acquisition and maintenance of at least some of our implicit biases. I then survey a response to Williams (1973) put forward by Hieronymi in her (2008) account of answerability, and consider whether Hieronymi's account reinstates the substantial distinction (§5.2.3). I suggest that in fact Hieronymi's account favours the continuum theory. I argue that if Hieronymi's account was to succeed in showing that we lack direct control of the acquisition and maintenance of our implicit biases, it would thereby also show that we lack direct control of the of the acquisition and maintenance of at least some of our beliefs. Because of this, the argument that there is a substantial distinction between the kind of control that we exert over the acquisition of our beliefs as compared with the kind of control that we exert over the acquisition of our implicit biases, fails.

5.2.1. *A provisional victory for the continuum thesis?*

If one's chosen theory of control is that of *voluntary* control, then it would seem that we lack this with respect to the acquisition of our beliefs, as many have argued (Williams, 1973; Bennett, 1990; Feldman, 2001; Strawson, 2003; Hieronymi, 2008). Williams (1973) maintains that we cannot simply believe at will, that is, believe whatever we want, for the very reason that we take our beliefs to represent reality. He argues

> If I could acquire a belief at will, I could acquire it whether it was true or not. If in full consciousness I could will to acquire a 'belief' irrespective of its truth, it is unclear that before the event I could seriously think of it as a belief, i.e. as something purporting to represent reality. (Williams, 1970: 148)

For instance, I cannot believe at will that my living room walls—which, as I look at them now, appear to me to be white—are, in fact, orange. I can *imagine* that my living room walls have turned orange, and I can *wish* that someone would come in and decorate the room, painting them orange. However, if Williams is correct, then as long as I have no evidence that my walls are orange, it is not within my

direct control to bring myself to believe that they are just by wanting to believe it.[47]

Others (Strawson, 2003; Levy, 2005) appeal to the phenomenology of acquiring a belief to demonstrate that belief acquisition is not voluntary. Strawson maintains that:

> …the role of genuine action in thought is at best indirect. It is entirely prefatory, it is essentially—merely—catalytic. For what actually happens, when one wants to think about some issue or work something out? If the issue is a difficult one, then there may well be a distinct, and distinctive, phenomenon of setting one's mind at the problem.... No doubt there are other such preparatory, ground-setting, tuning, retuning, shepherding, active moves or intention initiations. The rest is waiting, seeing if anything happens, waiting for content to come to mind.... There is I believe no action at all in reasoning and judging considered independently of the preparatory, catalytic phenomena just mentioned, considered in respect of their being a matter of specific content-production or of inferential moves between particular contents. (Strawson, 2003: 231-3; quoted in Boyle, 2009: 133)

The idea is that whilst agents may well bring particular thoughts to mind which have a bearing on the issue of whether to accept a new belief that *p*, when it comes to actually forming the belief, the agent must simply wait and see which new beliefs happen to bubble up in the mind. If one is convinced by Williams

---

[47] One might recall the examples of *Self-doubt, Grief,* and *Explicit Prejudice* from the previous chapter, in which agents fail to form beliefs on the basis of the evidence which they take themselves to have. It might be thought that these agents are doing something *voluntary* in failing to believe in line with their evidence, and that, therefore, these agents violate Williams' (1973) contention that belief acquisition is involuntary. But this is not the case. For whilst the agents in *Self-doubt, Grief,* and *Explicit Prejudice* violate Gendler's (2008a, 2008b) claim that recognition of evidence that *P* is *sufficient* for belief that *P*, they do not violate Williams' claim that recognition of evidence that *P* is *necessary* for belief that *P*. All these agents have (what they take to be) evidence for the beliefs which they fail to update in light of further evidence. For example, recall that:

> Ada (mistakenly) recognises Luke and Steven's laughter at her presentation as evidence that she is bad at maths.
> Ada forms the belief that she is bad at maths.
> Ada recognises her high test scores as evidence that she is good at maths.
> Ada fails to acquire the belief that she is good at maths.

It's not the case that Ada (thinks she) has no evidence that she is bad at maths—she interprets Luke and Steven's laughter as evidence that she is bad at maths. As she does not believe that she is bad at maths on the basis of having no evidence, she does not count as believing at will.

(1973) and Strawson (2003) and the rest quoted above, then one may advance the following continuum claim: whilst implicit biases are not acquired in an exercise of voluntary control, neither may beliefs be acquired in this manner, and so that we lack voluntary control over our implicit biases does not distinguish them from beliefs.

5.2.2. *SD argument from indirect doxastic control*

The above continuum claim is premature, however, because all that the particular argument from Williams (1973) shows is that we lack *direct* voluntary doxastic control. This result is compatible with us exercising *indirect* voluntary doxastic control. We said that one has indirect control of one's ϕ-ing, if one can directly ψ, and thus bring it about that one thereby ϕ-s. Indeed, we do seem to have this sort of control over our beliefs. I can generate new beliefs, or update my current beliefs, by influencing my environment so as to make it the case that I have evidence for these new beliefs. In the oft used example, I can acquire the belief that the lights are on in an exercise of indirect voluntary control: *step 1,* I get up and turn on the lights; *step 2,* I look, and acquire the belief that the lights are on. We regularly exercise indirect voluntary control over our beliefs in less mundane ways than this. For instance, a person might want to regularly acquire new beliefs about recent current affairs, and so, in an exercise of indirect voluntary control, expose herself to news sources each morning.[48] Or, imagine a person who, when in conversation with his friends, realises that he has an inaccurate understanding of a particular topic in history. In an exercise of indirect voluntary control, he reads some library books on the topic in question, in order to update his erroneous beliefs with factually correct ones.[49] So, even if we cannot acquire beliefs in acts of direct voluntary control, we regularly exercise indirect voluntary control in order to acquire particular beliefs. That is, we can perform an act over which we have direct voluntary control (such as reading a newspaper or a library book) in

[48] To be precise, this is an example of deliberative, indirect, initiation control. That is, one's reading the news is the outcome of a deliberative process in which one decides to find out what is going on in the world. I think this is an example of initiation control, rather than intervention control, because the agent is acquiring a *new* mental state.

[49] I think that this agent is exercising both deliberative, indirect, *intervention* control (with respect to terminating some of his false beliefs) and deliberative, indirect *initiation* control, with respect to acquiring new beliefs.

which we expose ourselves to evidence, and as a result of doing so we can acquire new beliefs, or update existing beliefs.

If it is the case that we do not have this sort of indirect voluntary control over our implicit biases, then there is grounds for a substantial distinction claim on the basis of this lack of indirect voluntary control. As many have suggested (Holroyd, 2012; Levy, 2014a; Holroyd and Kelly, 2016; Mandelbaum, forthcoming), however, we do have indirect control over at least some of our implicit biases. Consider the following findings.

> *(i) Exposure to counter-stereotypical exemplars*
> Dasgupta and Greenwald (2001) demonstrate that participants who are exposed to counter-stereotypical exemplars (in particular, exposed to pictures of admired black celebrities, and well known, but disliked white individuals) manifest less race bias. This effect was shown not just shortly after exposure, but was also present 24 hours later.
>
> *(ii) Imagining counter-stereotypical exemplars*
> Blair, Ma and Lenton (2001) reveal that entertaining counter-stereotypical mental imagery reduced the manifestation of implicit bias on a number of psychological measures. Participants who spent a few minutes imagining "what a strong woman is like, why she is considered strong, what she is capable of doing, and what kind of hobbies and activities she enjoys" manifested less implicit bias than those in the control condition, whilst in a further experiment, those who imagined a stereotypically 'weak' woman manifested more implicit bias than the control group, (2001: 830). As researchers tested participants on a number of measures, they were able to determine that, in particular, imagining a strong woman allowed participants to control (or suppress) their implicit stereotypes of women specifically (2001: 837).[50]
>
> *(iii) Reading a strong argument*
> Subjects of Briñol et al. (2008; cited in Mandelbaum, forthcoming) who are presented with a strong argument for hiring an African American professor

[50] Recall from Chapter 1 that if we see a reduction in stereotypic responses on the IAT, we are not able to determine whether this is mediated by an endorsement of 'women' + 'strong', or an endorsement of not-'men' + 'strong'. However, the Go/No Go Association Test (GNAT), one of the measures employed by Blair et al. (2001) does allow us to determine which attitude mediates the change.

exhibit less bias than that exhibited by those presented with a weak argument (as already noted in Chapter 4, §4.2).

We said that one has indirect control of one's ϕ-ing, if one can directly ψ, and thus bring it about that one thereby ϕ-s. In each of the above cases, agents have direct control over ψ-ing, where *ψ-ing* in each case is (i) exposing oneself to counter-stereotypical exemplars; (ii) imagining counter-stereotypical exemplars; or (iii) reading a strong argument. The effect that agents bring about with their ψ-ing is a modulation of their implicit biases. So, insofar as agents are able to bring about the above changes in their implicit attitudes, they have indirect control of the maintenance of implicit biases (and the acquisition of new implicit attitudes).
I now consider two possible objections to the claim that we have indirect control over the modulation of implicit biases, and the acquisition of new implicit attitudes: (i) the epistemic conditions objection; (ii) the state change *vs*. bypass objection.

(i) Epistemic conditions objection
SD theorists might point out that the epistemic conditions for this sort of indirect, intervention control are relatively demanding, and require inferential awareness of at least some of the empirical findings on implicit bias. To act *because* of the recommendations of an empirical study, one has to be familiar with the empirical study in question. I think that continuum theorists can accept that one must meet these epistemic conditions in order to employ the control strategies as they are recommended by the empirical studies above, and be successful in reducing their biases. This alone refutes Saul's claim in JS3.[51]

I think that at least sometimes, people employ similar strategies to those above with the intention of forming or updating their attitudes, even when they are unaware of implicit biases and the research on indirect control strategies. For instance, a person may decide to stop reading a politically motivated newspaper which both publishes a lot of material on successful white men and seeks out stories on crime perpetrated by black people, and decide to do so because the paper's editorial motive, and its likely effects on her attitudes, makes her uneasy.

[51] JS3: Inferential awareness that we are likely to be implicitly biased is not sufficient for control over implicit biases (interpreted as implicit biases *qua* attitudes)

She might instead seek to get her news from a more balanced publication which runs stories on famous people of colour, as well as taking a critical stance towards the misdeeds of white public figures. Effectively, this agent is exposing themselves to counter-stereotypical examples, with the intention of updating their attitudes.

In light of this discussion, it is interesting to consider Saul's (2013) claim that the acquisition of implicit biases result solely from our culture.[52] Presumably by 'culture', she means cultural stereotypes in an agent's environment. Arguably, our explicit social attitudes result from our culture, but to the extent that we are able to exert control over at least some of the media to which we expose ourselves, we are able to exert (indirect) control over our cultural attitudes. As I have already mentioned, one can choose which news outlets to expose oneself to, which literature to read, and so on. This is even easier with the advent of technology where one can set things up such that the media that one receives is automatically filtered (consider a tailoring a Twitter feed, or the inputs on a personalised news homepage), and so it is quite possible to exert direct control over a great deal of content to which one is exposed by choosing to only browse content from reputable platforms which do not (actively) utilise harmful stereotypes. So this claim does not cut a sharp distinction between our implicit and our explicit cultural attitudes. Both originate in our culture, and we are able to exert indirect control over both, at least to some extent. So, just as we can exert indirect control over belief acquisition and adjustment, so too can we exert indirect control over at least some of our implicit biases.

(ii) State change vs. bypass objection

SD theorists may question how we know that these techniques bring about a genuine change in the implicitly biased attitude itself, rather than simply enabling agents to *bypass* the original attitude, restricting it from manifesting on attitude measures. We can reply by pointing out that in at least some contexts, the attitude change remains stable over an extended period—it was observed 24 hours later in the case of Dasgupta and Greenwald's participants (2001). But I think there is a more powerful reply to the SD theorist if we consider the bypass question when it

[52] This is the claim that I summarised in JS2, accordingly "We are not able to exert control over the acquisition of our implicit biases, in virtue of the fact that they result solely from our living in a bigoted culture."

comes to *beliefs*. When a subject first acquires a new belief that contradicts the content of another belief that they have long held, do they determinately eradicate that old belief, or do they merely 'bypass it'? I think that there are at least some cases of belief update where the bypass hypothesis is in fact the only hypothesis that explains the data.

Consider the example from the previous chapter in which a person learns that her belief that her friend lives on College Road is in fact false, and that her friend actually lives on Station Road. This person meets many of the criteria for having acquired a new belief: she assents to the proposition that her friend lives on Station Road, she informs others that this is the case, and this proposition figures in her reasoning as regards how to get to her friend's house on a number of occasions. Nevertheless, a few weeks later, she tells another friend that her first friend lives on College Road, fully believing herself to have stated something true. That her first friend informed her that they in fact live on Station Road has simply slipped her mind. Now, it would seem that the only way that she would be able to retrieve the information that her friend lives on College Road, even after learning that her friend in fact lives on Station Road, would be if she retained the old belief that her friend lives on College Road after all, and was simply bypassing it in previous contexts. So, in reply to the SD theorist, it is not clear that when we acquire a new belief that not-*P* after having previously believed P, we determinately overwrite *P* with not-*P*, rather than simply bypassing *P* when we utilise not-*P* in appropriate circumstances. So, the bypass objection is not problematic for the continuum theorist: even if we bypass old implicit attitudes rather than update them, we also do this, at least sometimes, in the case of beliefs.

So, there is no substantial distinction between the relevant attitudes on the basis of indirect voluntary control. We may have this kind of control over the acquisition and maintenance of both (i) agential attitudes such as beliefs, *and* (ii) implicit biases.

5.2.3. *SD arguments from direct doxastic control*

Other philosophers disagree with Williams (1973), Strawson (2003), and the rest who argue that control of belief is at best indirect, and maintain that there is a sense in which belief *is* directly controllable, even if the control is not necessarily voluntary. I will now briefly outline a prominent account of direct doxastic control: that of 'answerability' from Pamela Hieronymi (2008). I argue that in

order to show that we are answerable for *all* of our beliefs, the SD theorist must also accept that we will sometimes be answerable for our implicit biases. Therefore, the SD claim on the basis of direct doxastic control, alike the SD claim on the basis of indirect doxastic control before it, fails.

Like Williams (1973), Hieronymi (2008) maintains that beliefs are not under our direct *voluntary* control. She argues that we do, nonetheless, exercise a kind of direct control over them. To set up the argument, she compares beliefs to Anscombe's (1957) account of actions. Here, Anscombe holds that we are 'answerable' for our agential actions in the sense that it is appropriate to ask us to answer the question why we acted with a specific kind of reason: Such a question invites the person to give an account of the factors which they saw to favour acting. That is, it seeks to discover an agent's reasons for acting. Hieronymi (2008) holds that we may ask similar questions of believing agents—we may ask them to justify *why* they believe that *P*. Accordingly, she says:

> …whenever one believes that *p*…one can rightly be asked, "Why do you believe *p*?" where that question looks, not for an explanation of how it came about that one believes, but rather for considerations that one takes to bear positively on whether *p* (that is, roughly, one's reasons for believing). (Hieronymi, 2008: 359)

Compare this to the phenomena that Strawson was concerned with. Whilst Strawson's account of "…waiting, seeing if anything happens, waiting for content to come to mind" (Strawson, 2003: 232) may say something accurate about "how it came about that one believes [that *P*]", it is not an appropriate answer to the question of "*Why* do you believe that *P*?" However, it is entirely appropriate to ask Strawsonian and Williamsian believers to provide an answer to the second kind of question. The reason for which I believe that the living room walls are white, for instance, is that they appear to me to be white. The same is true for our evaluative attitudes. Whether or not I exercise voluntary control over my evaluation of Beethoven's piano sonatas as richer, more complex and more moving than Mozart's piano sonatas, it is appropriate to ask me to justify my evaluation by providing the reasons for which I made it.

Hieronymi argues that even though we do not have direct *voluntary* control over what we believe, we are not thereby passive with respect to our

beliefs (2008: 338). Rather, we are vested in our beliefs in virtue of the reasons for which we take them to be true. I am vested in my evaluation of Beethoven—and in all of my evaluations, whether they are aesthetic, prudential, political, moral, and so on. Hieronymi thus suggests that beliefs are "commitment-constituted attitudes" and argues that we have a distinctive form of (direct) control over them:

> Because these attitudes embody our take on the world, on what is or is not true or important or worthwhile in it, we control them by thinking about the world, about what is or is not true or important or worthwhile in it. Because our minds change as our take on the world changes…we can be said to be "in control" of our commitment-constituted attitudes. (Hieronymi, 2008: 370-1)

As I mentioned in §5.1.2, I take Hieronymi to be arguing here for a deep self account of doxastic control, because, in believing what is good and what is valuable, for instance, the agent expresses her fundamental evaluations of the world. (Although, as I also mentioned in §5.1.2, the analysis of what a fundamental evaluation turns out to be relies on the notion that it is appropriate to ask the agent to give *reasons* for believing. So this is a deep self account that analyses our fundamental evaluative stance in terms of our capacities as reasoners). By Hieronymi's own admission, this is also a theory of *direct* control (2008: 357). Believing is the agential act—the agent does not have to perform any prior ψ-ing in order to believe such that she is answerable for doing so.

If it can be shown that implicit biases do not embody our take on the world, and that we are not answerable for them, then there might be grounds for a substantial distinction between the former and beliefs. The SD theorist may argue that because implicitly biased agents do not consciously endorse any particular considerations *as* reasons for harbouring their implicit biases, therefore they are not able to answer "Why do you think that *P*?" (where *P* is the contents of their implicit bias) and hence they are not answerable for it. Because of this, it might be argued that implicit biases do not embody an implicitly biased agent's take on the world.

I don't think that this is quite right, however. Recall that Hieronymi's (2008) condition is not that agents *can* in fact answer such a question, but, given that they are somewhat vested in an attitude as it guides speech acts and other

behaviours, it is *appropriate to ask* them for their reasons. I think that there is a sense in which implicitly biased agents are answerable for particular utterances which reveal their implicit biases, even though they do not consciously affirm the considerations on which they acquired their implicit biases as reasons for doing so. Consider the following case:

> *Courier*
> A courier delivers a letter addressed to Dr Dewan, and when a woman answers the door, he asks if her husband is available to sign for his letter. His belief that the woman at the door is not Dr Dewan is a manifestation of an implicit bias which couples men and academic achievement, and women and homemaking. In fact, the woman at the door is Dr Dewan. She has won a prestigious research grant as is at home researching for her next book.

The courier has a belief which manifests a coupling of men and academia (or perhaps, women and being at home) even if only by implication. The courier does not act autonomically when he asks if the woman's husband is available. His action is guided by an occurrently tokened belief that the woman at the door is not Dr Dewan. The belief that the woman at the door is not Dr Dewan *embodies his take on the world*, as Hieronymi would say. So, even though he may well not affirm the coupling of men and academic achievement as a reason to assume that the woman at the door is not Dr Dewan if he was so asked, he nevertheless *does* believe that the woman at the door is not Dr Dewan. It seems entirely appropriate in this situation for Dr Dewan herself to ask the courier why he believes that she is not Dr Dewan. So, even if the courier cannot provide the reason "I have an implicit bias against women in academia" for example, as a reason for his belief, his belief does embody his take on Dr Dewan, and for this he seems fully answerable in Hieronymi's (2008) sense of the word.

Consciously affirming some considerations as reasons for holding an attitude is not the only way to endorse an attitude such that it embodies one's take on the world. By committing to the claim that one must consciously affirm the considerations for which one acquired an attitude as reasons for doing so, in order to be answerable for a particular attitude, the SD theorist will end up excluding at least some *agential* attitudes from the set of attitudes for which we are answerable. Attitudes may be formed without introspective awareness, on the basis of considerations which the agent in question does not consciously affirm as

reasons at the time the attitude is acquired. Nevertheless, these attitudes may go on to guide what looks very much like agential behaviour. Recall the three cases from Chapter 3, §3.2.3, where agents form and act on beliefs without occurrent introspective awareness: Muhammed acts on a belief about where the torch is kept, even though he is not occurrently introspectively aware of forming, or acting on, this belief; Laura acts on a belief about where Aisha's flat is, even though she is not occurrently introspectively aware of forming, or of acting on, this belief (at least in the instance of utterance, though of course having spoken as she does, Laura brings her belief to introspective awareness); and Naveen acts on a belief about the colour of the mushroom punnet, even though he is not occurrently introspectively aware of forming, or acting on, this belief in the moment of utterance. As such, neither Muhammed, Laura nor Naveen consciously affirm any considerations as reasons for forming their attitudes. And yet, these attitudes seem to have been formed in response to features of the situation which make them *reasonable*. For instance, Muhammed's unconscious attitude has the following sort of content "the cupboard door handle is such-and-such a distance in such-and-such a direction from my body"—content which is reasonable in light of the relative position of the cupboard to him. Further, there is a sense in which these agents' attitudes embody their take on the world. Muhammed takes it that the world is such that the cupboard door is at *this* angle, relative to his body; Laura takes it that the world is such that Aisha lives on the second floor, and Naveen takes it that the world is such that the mushrooms in the fridge are in a black punnet.

There are other cases where agents acquire attitudes without consciously affirming any considerations as reasons for their acquisition, but which nevertheless appear to make some forms of agential behaviour possible. Angela Smith (manuscript: 17-18) appeals to the notion of 'flow', a psychological state observed by Csikszentmihalyi (1990) and Gladwell (2005) in which agents fully immersed in an activity do not have introspective awareness of the mental states guiding their actions. For instance, a skilled jazz musician may improvise novel sequences without effort or awareness of the processes which guide her actions. All the time she is responding to cues from her fellow musicians, thus she is forming representational mental states which guide action. As Smith suggests "Her musical reactions are clearly *reasons-responsive*—she is exquisitely attuned and responsive to the playing of her fellow musicians—but this reasons-

responsiveness operates below the level of conscious awareness" (manuscript: 17-18). The musician is also unaware of these representational states as they are acquired, and so does not *consciously* affirm the considerations on which she acquired these states as reasons for doing so. The musician cannot answer the question of "Why do you think that *P*?" where *P* is any of the representational states that inform her playing, in a sentence. Nevertheless, because such states inform her playing, there is a sense in which she endorses them, in their guidance of her responses. As they inform her musical responses, they embody her take on the improvisation session.

If the attitudes which count as "embody[ing] our take on the world, on what is or is not true or important or worthwhile in it" (Hieronymi, 2008: 370) must be consciously affirmed, and if the reasons for which an agent holds an attitude must also be consciously affirmed, then Muhammed, Laura, Naveen, and agents experiencing flow, end up as not answerable for their attitudes, even though, as I have argued previously, these attitudes are rightly considered as agential. In order to account for the cases of Muhammed, Laura, Naveen, and flow as agential, I think that we have to reject a substantial distinction account on which conscious affirmation of the considerations for which one holds an attitude as reasons for doing so is a necessary condition for answerability. If that is so, the implicitly biased agents such as the courier *do* count as answerable for their attitudes. So, if Hieronymi's (2008) account of answerability shows that agents directly control *all* of their beliefs, then is also shows that agents directly control at least some implicit biases, and the substantial distinction argument fails.

Let us take stock. If the foregoing is correct, then it would seem to be that there is no substantial distinction between the control that we exert over the acquisition of agential attitudes, such as beliefs, and that which we exert over the acquisition of implicit biases. I showed this to be the case both for accounts of indirect voluntary control, as well as for direct doxastic control. If control over the acquisition of implicitly biased attitudes doesn't distinguish them from beliefs, perhaps control of associated actions does. It is to this question that I now turn.

5.3. Control of Implicitly Biased Actions

Implicit biases typically affect behaviour automatically. That is, they become primed and ready to manifest in behaviour, or in fact do manifest in behaviour,

without our explicit intentions that this be so. This means that control efforts will often have to be focused on *intervening* on this process of automatic activation and manifestation. However, this does not mean that we do not exert *direct* control over such behaviour in at least some instances—as I argued in §5.1.2, the direct/indirect distinction cuts across the initiation/intervention distinction. Accordingly, an agent has direct control of ϕ-ing if it is within their power to ϕ, without any intermediary steps. An agent has indirect control of ϕ-ing if it is within their power to ψ (through an exercise of *direct* control, as above), and from the occurrence of ψ, the occurrence of ϕ follows. Agents have initiation control when they initiate an action sequence, and intervention control when they terminate or re-direct an action sequence that is already in progress.

In this section, I demonstrate that there are three different control strategies which are effective over at least some implicitly biased actions, which are also necessary for controlling at least some of our everyday agential actions. This undermines the SD theorist's claim that there is a kind of control that we have over all of our belief-guided actions which we do not have over our implicitly biased actions. I start by presenting a kind of indirect, intervention control that we have over many of our implicitly biased actions (§5.3.1). This indirect, intervention control is the *only* kind of control available to us to guide and hone many of our uncontroversially agential actions. I then argue (in §5.3.2) that agents may exercise two kinds of direct, intervention control over the manifestation of implicit bias in action: (i) a form of deliberative, direct, intervention control, and (ii) a form of non-deliberative, direct, intervention control. I demonstrate that we rely on both (i) and (ii) in everyday agential action.

Of course, SD theorists about control need it to be the case that there is a kind of control that we exert over all of our agential actions which we do not exert over any of our implicitly biased actions. In showing that there are many agential actions which are *only* controllable via strategies which we can *also* use to control implicitly biased actions, it follows that the SD theory as regards control of action is false.

5.3.1. *Indirect, intervention control of implicitly biased action*

In §5.1.2 we saw that there are some automatic actions on which we can directly intervene. For instance, an individual can directly intervene on actions such as nail-biting by directly ceasing to bite their nails. According to Holroyd and Kelly

(2016) agents *cannot* directly intervene on the manifestation of implicit bias in action:

> Recall Snow's idea …[of] intervention control—the ability to exert influence on autonomously running processes by stopping them or redirecting how they shape action. According to this understanding of intervention control, individuals lack it in relation to implicit biases—it is very difficult to prevent behavioural manifestation of implicit bias via direct reflective control. The job interview panellist cannot effectively intervene on the operation of implicit biases as they influence cognition, simply by thinking: 'Oops, there it goes; better get my cognitive processes back on track and stop that biased evaluation.' (Holroyd and Kelly, 2016: 127)

However, Holroyd and Kelly develop an account on which agents can control the manifestation of bias in their actions by *indirectly* intervening on these processes. With this notion of indirect, intervention control in play, they maintain that

> An agent can intervene in some automatic process [that would otherwise manifest implicit bias] not by bringing it under direct reflective control at the moment of its activation, but by diverting its activation by means of some environmental or cognitive prop put in place to derail unwanted cognitive or behavioural patterns. (Holroyd and Kelly, 2016: 127)

On Holroyd and Kelly's account, agents cannot simply ϕ, and in ϕ-ing, effectively intervene on the manifestation of implicit bias in action. However, they can (directly) manipulate either features of their environment, or features of their own cognitive processes, where these manipulations have the effect of intervening on the manifestation of implicit bias in action. As such, Holroyd and Kelly present an account of indirect, intervention control of implicit bias.[53]

---

[53] Holroyd and Kelly (2016) present three ways of deploying this indirect, intervention control: (i) by way of environmental props consciously employed for guiding cognitive processes; (ii) with cognitive props consciously employed for guiding cognitive processes; and (iii) through automatic processes as props unconsciously employed for guiding cognitive processes. I talk about (i) and (ii) in this section, but I disagree that the example that Holroyd and Kelly employ to demonstrate (iii) really is an example of indirect, intervention control, and instead suggest that it is better characterised as *direct*, intervention control, albeit of a non-deliberative kind. For this reason, I will postpone the discussion of (iii) until §5.3.2, where I discuss direct, intervention control.

Holroyd and Kelly present some of the environmental change strategies that I considered in §5.2.1, in which it was shown that people are able to update or reduce the strength of their implicitly biased attitudes (Dasgupta and Greenwald, 2001; Blair, Ma and Lenton, 2001). One of the effects of acting to alter our implicitly biased attitudes is a reduction of their manifestation in our behaviour, thus 'nipping' the origins of implicitly biased action 'in the bud'. Holroyd and Kelly maintain that

> …a person might rein in the expression of her own implicit racial biases by putting up pictures of admired black celebrities around her office, thus taking indirect, ecological control over those biases so that her judgements and actions more fluidly express her character and values. (2016: 122)

Holroyd and Kelly also consider what they call 'cognitive props' which can be employed by the agent to exert indirect, intervention control over the manifestation of an implicit bias, (2016: 122). Research shows that people have indirect control over the manifestation of their implicit biases in behaviour, by deploying 'implementation intentions' (Webb, Sheeran and Pepper, 2012). Implementation intentions are intentions with a conditional structure in which the subject plans to think about something, or to carry out an action, when they encounter a target concept (Stewart and Payne, 2008; Webb and Sheeran, 2008; Mendoza, 2010). In the case of experiments which seek to determine the effects of implementation intentions on implicit bias, such intentions tend to take the following form: "If I encounter [target person x] then I will think [social attribute *y*]" where *x* and *y* constitute a counter-stereotypical pairing. To measure the effects of such intentions, typically subjects first undertake tasks which serve as measures of their base rate level of implicit bias (an IAT or other implicit measure). Then subjects are instructed to form the relevant implementation intention, before they undergo another IAT test. For an example of how this instruction is given, subjects in Webb, Sheeran and Pepper's (2012) study saw the following onscreen message:

> Most people associate females with liberal arts and males with science subjects. Your goal in the following experiment is NOT to stereotype women. (Webb, Sheeran and Pepper, 2012: 17-18)

and were then asked to form the following implementation intention:

> If female and science are paired at the top of the screen, then I will respond especially fast to both science and female words! (Webb, Sheeran and Pepper, 2012: 18)

It turned out that participants who formed this implementation intention responded more quickly on stereotype-incongruent trials than those who did not form such intentions, revealing that such a strategy can enable participants to modify the extent to which their implicit biases manifested in behaviour (Webb, Sheeran and Pepper, 2012).[54] Employing implementation intentions in this way, subjects have indirect, intervention control over their implicitly biased behavioural responses. In an act of direct control, they call to mind a non-stereotypical concept, and this has the effect of mediating the implicitly biased response. This is a form of indirect *intervention* control, because the act of calling to mind a non-stereotypical concept intervenes on the activation of the implicit attitude which would otherwise produce a (more) biased behavioural response.

SD theorists might point out that employing implementation intentions to control our behaviour is a relatively demanding method of action control, and is quite unlike the control that we exert over our belief guided actions. Holroyd and Kelly (2016) disagree, arguing that implementation intentions are very much a part of our everyday agential control. They maintain that agents utilise strategies similar to implementation intentions to hone and refine various everyday skills (2016: 122). In their discussion of implementation intentions as a strategy for taking indirect, intervention control over the manifestation of implicit bias, they draw an analogy to a sports player who

---

[54] You may recall that whilst the IAT measures stereotype matching behaviour, it is usually taken to reveal one's underlying (implicit) attitudes, because it requires fast responses which may not be altered by general conscious intentions to respond non-stereotypically. However, this does not rule out that IAT performance may be altered by something other than these general conscious intentions—in which case, modulated scores on the IAT may not necessarily indicate altered underlying attitudes, if one has reason for thinking that another variable may account for modulated responses. Webb and colleagues' claim that the present study does not reveal a change in the underlying attitudes, but rather a change in the control that one exerts over the manifestion of attitudes in behavior (2012: 16) relies on previous findings which show that implementation intentions do not necessarily change underlying attitudes, but rather enable the agent to control the extent to which these affect behaviour, using a number of different paradigms to the IAT (for discussion, see Webb and Sheeran, 2008; also, Gollwitzer and Sheeran, 2006; Webb and Sheeran, 2006).

> …practises, calibrating the operation of sub-personal subsystems to bring them in line with intentions, and thus developing a certain kind of fluid and unthinking control. (2016, 122)

Indeed, it would seem that utilising implementation intentions is a necessary part of the process of enhancing fine motor control.

This sort of control seems uncontroversially agential. As Arpaly has argued, even though sports players employ various indirect control strategies to hone their techniques, it would seem incorrect to claim that therefore they do not act agentially when performing actions honed in this manner (2003: 52). Furthermore, I think that implementations intentions (or something very similar to them) are useful for enhancing more than just fine motor control, and are likely to be utilised by all sorts of people, in all sorts of agential pursuits. For instance, an actor might employ implementation intentions to develop a particular expression (reflecting on a sad episode in his life to enhance his expression of emotion). A trainee accident and emergency doctor, anticipating the arrival of some car crash victims, might use implementation intentions to steel herself to maintain a sense of calm and professionalism in the face of appalling trauma. In particular, the doctor utilises indirect, *intervention* control, intervening on the manifestation of anxiety in her behaviour, and either suppressing anxious behaviour that is already in process, or preventing such behaviour from manifesting in the first place. It would seem inappropriate to claim that the actor's expression and the doctor's professionalism are not agential simply because they were enabled by implementation intentions. As Holroyd and Kelly point out, this type of control is in fact rather mundane, and "underlies a vast swathe of human behaviour and problem-solving" Holroyd and Kelly (2016: 123). So, because there are at least some agential actions, the performance of which would seem to require indirect control strategies, there is no grounds for a substantial distinction between *all* agential actions and all implicitly biased actions on the basis of the kind of control that we have of each.

I think that we are now in a position to refute the following two SD claims:

K&R4: Implicit biases influence action automatically.

JS3: Inferential awareness that we are likely to be implicitly biased is not sufficient for control over implicit biases (interpreted as implicitly biased actions)

Regarding K&R4, whilst it may be the case that implicit biases influence action automatically, this does not rule our being able to intervene on their operation. Similarly, the trainee doctor's anxiety may influence her action automatically, but this does not mean that she cannot take steps to prevent it from manifesting. So, it does not follow from K&R4 that we have any less control over the manifestation of implicit bias in action than we do over a number of agential actions. Regarding JS3, inferential awareness of implementation intention studies, for instance, *is,* at least sometimes, sufficient for (indirect, intervention) control over the manifestation of bias in action. So this claim will not uphold a substantial distinction either.

The SD theorist might raise an objection about the epistemic conditions for utilising these indirect, intervention strategies to control the manifestation of implicit bias in behaviour, similar to that discussed in §5.2.1. The objection goes like this: such strategies would seem to require knowing about findings from cognitive science, and so, it might seem that agents can only utilise these indirect, intervention control strategies if they are aware of the relevant findings. Further, the SD theorist can point out that we don't have to know about empirical findings to utilise indirect, intervention control in everyday cases of behaviour (such as those of the actor and the doctor). So, there would seem to be grounds for a distinction between control over implicitly biased actions and control over agential actions, with respect to the relevant epistemic requirements.

I resisted this claim with respect to attitude control in §5.2.2, and I resist it again here, with respect to action control. One does not need to have inferential awareness of implicit bias and the relevant findings in order to expose oneself to counter-stereotypical exemplars with the purpose of updating one's attitudes, and since a result of changing a biased attitude is that the attitude will no longer be available to guide behaviour, I think that the same can be said here: agents do not need to have inferential awareness of implicit bias in order to expose themselves to counter-stereotypical exemplars with the purpose of updating attitudes which might otherwise manifest in behaviour.

As for implementation intentions, I don't think that it is impossible that an agent with no inferential awareness of implicit bias might nevertheless still employ something *similar* to an implementation intention to control the manifestation of implicit bias in behaviour. Recall Borgoni's (2015) example of Emilia, who catches herself thinking implicitly biased thoughts about women in politics. Disturbed by such thoughts, Emilia might use something like an implementation intention to prevent such thoughts from entering any further into her deliberations. She might make the following sort of plan: "if I catch myself thinking that women are naturally less able in the field of politics than men, then I shall stop myself short and think instead of able political women that I know." This is to say that agents like Emilia, who are disturbed by their inegalitarian thoughts, are capable of making plans to try to prevent them from influencing deliberation any further, some of which might have the structure of implementation intentions. Whether or not such strategies in fact *are* successful at reducing biases in action is an empirical question. Until there is further research on such strategies, continuum theorists must concede that it remains the case that utilising the specific implementation intentions which *are* shown to be effective in current empirical studies requires inferential awareness of the studies in question.

But even if we make this concession, I think that the SD theorist still does not have sufficient grounds for a substantial distinction between the indirect, intervention control that we have over our implicitly biased actions and that which we have over agential actions, on the basis of the relevant epistemic preconditions. As I argued above, agents do not require inferential awareness of the studies on counter-stereotypical exemplars in order to expose themselves to inclusive media, with the effect of reducing the manifestation of bias in action.

Another point that is relevant to this discussion is that some *agential* activities, such as honing a particular skill in sport, might not be possible until one acquires inferential knowledge of particular developments in sport science. For instance, knowing the precise arc of motion, and the angle at which the arm delivers the most power, might be necessary for developing one's serve. And yet, a game-winningly powerful serve, even if it is developed and honed on the basis of learning about findings in musculoskeletal biomechanics seems to be an agential action nonetheless. We don't deprive athletes of titles on the basis that the development of their skills depends on their knowing particular findings in sport science.

So, to summarise my response to the epistemic conditions objection: in utilising the indirect, intervention control strategies mentioned above, agents sometimes require inferential awareness of the relevant scientific findings, and sometimes they don't—and this is true of both implicitly biased actions and of agential actions.

5.3.2. *Direct, intervention control of implicitly biased action*

Recall that an agent has direct control of some ϕ-ing if it is within their power to ϕ, without any intermediary steps. This is to say that the agent's bringing about of her ϕ-*ing* is itself the act of agential control. Recall also Holroyd and Kelly's contention from the last subsection, that it is very difficult to prevent the manifestation of implicit bias via direct control and that people "cannot effectively intervene on the operation of implicit biases as they influence cognition, simply by thinking: '…better get my cognitive processes back on track and stop that biased evaluation'" (Holroyd and Kelly, 2016: 127). Their claim is not that it is *impossible* to have any form of direct control over the manifestation of implicit bias, but that it is, at least, "very difficult".

In what follows, I outline some cases in which agents have what I think is properly characterised as direct, intervention control over the manifestation of implicit bias in action. Contra Holroyd and Kelly's claim, this is also control which does not come at a great effort. I argue that there are cases in which agents may ϕ, where ϕ-ing is both within their direct control, and where the ϕ-ing itself constitutes an intervention on the manifestation of an implicit bias in action. I argue that, at least sometimes, agents have direct intervention control over the manifestation of implicit bias in action. I first discuss a kind of (i) deliberative, direct, intervention control, and then a kind of (ii) non-deliberative, direct, intervention control.

(i) Deliberative, direct, intervention control of implicitly biased action

Sometimes, when an implicit bias manifests in action, the agent would seem to act in a way that they would not have done, had the implicit bias not been present, and had the action instead been guided wholly by the values and intentions that the agent professes to have. For instance, Henry's implicit bias against black people manifests in his blinking *more*, when in conversation with a black colleague than he otherwise would have, had he not had this bias. We can call this

a ‘manifestation case’. Other times, an implicit bias would seem to block an action from being performed in a particular way, and instead, the agent will perform that action in a restricted or limited way—a limitation that would not have been there had the implicit bias not manifested. For instance, Franz’s implicit bias against women manifests in his failing to invite women professionals to speak on the business development panels that he organises. In this case, Franz is performing an action (the selection of speakers) in a restricted manner, in that, whilst his intention is to select speakers from all areas on the basis of their expertise, he ends up primarily selecting questions from male business experts, and so he is performing this selection action in a manner that is restricted, relative to his intentions. Call this a ‘restriction case’.

It may be that some restriction cases can be re-described as manifestation cases, and *vice versa*. For instance, consider Olivia’s implicit bias against Muslims, which manifests in her crossing the street to avoid a Muslim man. We could describe Olivia as (1) “walking in a manner which decreases proximity with a Muslim”. Because this description picks out the way in which the implicit bias manifests in the current action, it would seem to be a manifestation case. But equally, we might want to describe Olivia as (2) “failing to perform an act of tolerance”. Because this description picks out a failure of action performance, it would seem to be a restriction case. That some restriction cases may be re-described as manifestation cases, and *vice versa*, isn’t a problem for what I want to say. For my upcoming argument to work, I only require that there are at least some cases of implicit bias where it is possible for the agent to *realise* that they are performing an action in a way that is restricted *relative to the way that they intended* to perform that action, as well as to realise that it is possible for them to perform less restricted actions from then on. As such I am interested in manifestation *vs.* restriction as an epistemological distinction, whether or not there is a robust metaphysical basis to the distinction.

I propose that, in restriction cases, if the agent can (i) realise that a biased preference has blocked them from performing an action in such-and-such a way, and (ii) it is possible for them to deliberately and directly perform the action in that way from then on, then they may directly intervene on the manifestation of implicit bias in action. By ‘directly’ performing the intervention action, I mean that the agent ϕs, where their ϕ-ing either cuts off a current manifestation of

implicit bias in action, or prevents its continued manifestation. Call this '*un*restricting'.

Restriction cases, I contend, are relatively common. Further, they are common *both* for implicitly biased agents *and* for agents whose agential preferences manifest in action without their (initially) realising it. Both sorts of agents may directly intervene on the manifestation of the bias/preference in action in a number of these cases, thus 'unrestricting' non-biased/non-preferential actions which may be performed from then on. Let's look first at some unrestriction cases featuring implicitly biased actions.

> *Seminar:* A seminar chair's implicit bias manifests in his predominantly selecting questions from white men in the discussion session. He both can, and, in this instance (i) does notice that he predominantly selects questions from men, and (ii) takes direct control of the manifestation of his implicit bias, by intervening and *actively looking* for women and people of colour who have raised their hands.[55]
>
> *Author:* The implicit bias of a highly successful and well read author of fiction manifests in a number of novels where white, male characters dominate the dialogue and the plot. She both can, and, on this occasion (i) does notice that the majority of her main characters are white men, and (ii) takes direct control of the manifestation of her implicit bias, by intervening and *deliberately writing* in a more inclusive set of characters in her next novel.
>
> *Panel show producer:* A panel show producer's implicit bias manifests in the selection of participants who are almost always white men. He both can, and, in this instance (i) does notice this fact, and (ii) takes direct control of the manifestation of his implicit bias, by intervening and *actively seeking out* participants from more diverse backgrounds.

In each of the cases, participants gain awareness of the manifestation of implicit bias in action. This seems quite possible. All of these cases are cases in which

---

[55] I owe thanks to Jenny Saul for suggesting this particular example of an unrestriction case, during the Q&A of my presentation on control of implicit bias at the Leeds Minorities and Philosophy Conference on Implicit Bias, October, 2015.

participants have observable class preferences (Chapter 3). As I argued in Chapter 3, I think it is possible that, at least sometimes, people may discover some of their implicit biases through introspection on their particular preferences as regards social groups. However, observational awareness of the manifestation of a biased preference in behaviour is quite sufficient in the above cases for people to be able to directly perform the relevant intervention behaviours, whether or not they also have introspective awareness of their biased preferences.

Note that it is not necessary to have any inferential awareness of implicit bias and the relevant empirical findings to be able to make the sorts of observations necessary for redirecting action as above. Once people notice that a particular behaviour is shaped by a biased preference, they can deliberately redirect their action, thus 'unrestricting' the action they would have performed had their biased preference not manifested. The unrestricted action constitutes a direct intervention on the manifestation of implicit bias. So, unrestriction cases show that agents can, at least sometimes, have direct, intervention control over the manifestation of implicit bias in action, and, further, contra Holroyd and Kelly (2016: 127) this is relatively easy for the agents in question to do.

One might think that it is somewhat demanding to expect implicitly biased agents to notice biased behaviour, and to exert this sort of control. I am not convinced that it is. Many of our everyday preferences frequently guide behaviour without our express intention that they do so. A politics graduate may have an everyday preference for the left-wing of politics, for instance, which she endorses explicitly, and which she frequently employs in introspective practical reasoning to determine her actions (such as how to vote, for instance). However, this preference may also guide her behaviour without her express intention in many other contexts. For instance, she reads news articles which endorse the protection of state-funded institutions more frequently than those which support shrinking the state and placing public institutions in private hands, even though she doesn't expressly intend to do this. In many of the contexts in which her left-wing preference guides her actions automatically, were she to become aware (either introspectively or observationally) that this is the case, she would endorse its being so—as would be the case in the example of her news reading preferences. However, in other contexts, were she to become aware that her left-wing preference automatically guides action, she would not endorse its manifestation in that particular action. I will give examples of this shortly. I think there are plenty

of cases in which agents both can, and are in fact *expected* to both (i) notice that their preferences automatically guide their actions in a fashion inconsistent with their present aims, and (ii) to redirect their future behaviour in line with a non-preferential course of action. Consider the following everyday unblocking cases:

> *Student council:* The student council chair's left-wing preference manifests in the selection of questions from known left-wingers at a much higher rate than questions from known right-wingers during a council meeting, without her initially realising it. Given that her role requires that she ensures that people on either side of the political spectrum have an opportunity to have their say, she is both expected to, and, in this instance, does (i) eventually notice the manifestation of a left-wing preference in question selection, and (ii) takes direct control over this, by intervening and actively looking for questions from those she knows to be right-wing.
>
> *Reports:* A researcher repeatedly favours particular adjectives throughout his reports, without his initially realising this. Given that his role requires that he communicates concepts in an articulate and effective manner, he is both expected to, and, in this instance, does (i) eventually notice the manifestation of his linguistic preferences and (ii) takes direct control over them, by deliberately selecting a close synonym the next time he finds himself considering using one of his overused adjectives.
>
> *Debate:* A TV executive's advocacy of climate skepticism manifests in the programming of excessive airtime for climate skeptics on her debate show, without her initially realising this. As an executive for a publicly funded channel, she is both expected to, and, in this instance, does (i) eventually notice the manifestation of her preference for skepticism, and (ii) takes direct control over it, by actively inviting contributions from climate scientists.

For each agent in the above three examples, a preference which they may endorse in some contexts manifests automatically in a context in which, when they become either introspectively or observationally aware of it, they do *not* endorse its guidance of action. It seems quite possible for the agents above to become at least observationally, if not introspectively (or perhaps retrospectively) aware of how their preferences guide their actions. Furthermore, I think that, given the role that each of these agents is performing, they are expected to notice the

manifestation of preferential treatment in their behaviour, and to do something about it: the student council officer ought to make sure that she is facilitating a fair debate; the report writer ought to make sure that he is describing complex concepts in the most articulate and meaningful language; the TV executive for a publicly funded channel ought to be producing an unbiased show. So, even though action may often be guided by the automatic manifestation of preferences, it is possible for agents to become aware that this is the case, and to intervene on the manifestation of the preference, thus unrestricting an action that they *would* endorse in the context at hand. These agents have a deliberative form of direct, intervention control over the manifestation of their preferences in action, in that, on recognising the preference, they ϕ, where ϕ is actively looking for questions from right-wingers; deliberately selecting a close synonym; or actively inviting contributions from climate scientists, and where ϕ-ing constitutes a direct intervention on the manifestation of the preference in action.

So, there are at least some cases where everyday preferences affect action automatically, in a way that the agent in question would not endorse, but over which they have direct intervention control. When these everyday preferences manifest automatically, they aren't subject to initiation control, and so there will be times when the utilisation of intervention control is the *only* way that agents may control these actions (although they can deploy this intervention control directly). Not only is their doing so a belief-guided, agential action, it is also expected of them. So, we have another kind of control (direct, intervention control), in addition to that presented in the previous subsection (indirect, intervention control) that (i) is *necessary* for controlling at least some instances of everyday, belief-guided actions, and (ii) is also available to agents in the case of at least some implicitly biased actions (the unrestriction cases). As such, the SD claim that there is no strategy by which we may control at least some of our implicitly biased actions that it is also necessary to use to control at least some of our belief-guided actions, fails yet again.

What of Holroyd and Kelly's claim that a job interview panellist cannot directly intervene on the operation of their implicit biases, simply by thinking "Oops, there it goes; better get my cognitive processes back on track and stop that biased evaluation" (Holroyd and Kelly, 2016: 127)? When it comes to exerting control over an evaluation, in which the agent takes into account a candidate's various skills and experience, and in which they may also (unintentionally) take

into account the candidate's race and gender, I agree with Holroyd and Kelly (2016: 127) that it is not clear what an agent can do to "get her cognitive processes back on track". It isn't immediately obvious to the agent what she can do to intervene on the operation of the implicit bias, since it isn't obvious how the bias restricts her thinking. At least, it is not as obvious as the unrestriction cases that I discussed earlier, where there is a clear course of action that the manifestation of bias previously blocked the agent from taking, but which they can take from then on.

However, not knowing exactly how an implicit bias manifests in evaluation is not, I think, wholly a function of the *implicit* aspect of the attitude in question: It is also, in part, a function of the *biased* aspect of the attitude. The same problem arises when an *explicit* personal preference interacts with the evaluation of some objective criteria. For instance, I have an explicit preference for my friend's character and demeanour. Were I to be a on the panel for a job to which he applied, and were I to have to evaluate his suitability for the position compared with a number of other applicants, it is likely that my explicit preference for his character and demeanour would guide my evaluation, even if I did not intend for this to be the case. It is not clear that I would have any more insight into how my explicit preference for my friend guides my evaluation of his credentials than the insight that I have into how, for instance, my implicit race bias guides my evaluation of his credentials.

To prevent favouring my friend on the basis of my explicit preference, I could collate all of the objective information (exam results, relevant qualifications, and so on) that I can attain from the applicants' C.V.s anonymously, and resolve to select the candidate who is objectively best on paper for the job. Doing this would prevent the manifestation of my preference for my friend. However, as Holroyd and Kelly acknowledge (2016, 122), this option is open to us in the case of *implicit* bias as well—we can anonymise C.V.s (thus preventing gendered and racialised names from activating our implicit biases) and evaluate candidates purely on the basis of objective criteria.

Even here, though, it is not clear that I could block the manifestation of personal preferences completely. For instance, if I happen to prefer qualifications from institution *A* to qualifications from institution *B*, (and, suppose that I am unwarranted in doing so) then it might be that simply removing the name from a C.V.˙ is not sufficient to prevent the manifestation of my personal preferences in

my evaluation. I might say to myself "I won't let my preference for a qualification from *A* when this person only has a qualification from *B*, affect my overall evaluation of their suitability for the position." Whether I'll be successful in directly suppressing the manifestation of my preference for qualifications from institution *A* in my overall evaluation of a candidate is about as unclear as whether I'll be successful in directly suppressing the manifestation of a preference for, say, a white candidate, with an equivalent internal utterance of not letting race affect my overall evaluation.

So, when we're faced with with the task of evaluating candidates, we may not be able to exert direct, intervention control over the manifestation of our implicit biases. To take control, we may instead have to commit ourselves to the long-term, indirect intervention control strategies discussed by Holroyd and Kelly (2016), and which I discussed in the previous subsection. However, we are also unable to exert direct, intervention control over the manifestation of our *explicit* personal preferences when evaluating candidates. (This doesn't necessarily mean that we aren't doing something agential when either explicit or implicit preferences guide our evaluations—one might still think that these evaluations still "embody our take on the world" for instance. Rather, the point is that we cannot discern exactly how our preferences operate.)

So, to sum up this section, contra Holroyd and Kelly, I think that we *can* have a form of deliberate, direct, intervention control over our implicitly biased actions. Furthermore, there are cases in which this deliberate, direct, intervention control is *necessary* for controlling at least some of our everyday belief-guided actions. That at least some implicitly biased actions are amenable to the same kind of control as everyday belief-guided actions is inconsistent with the substantial distinction theory, and may only be accounted for on a model on which implicitly biased actions and belief-guided actions overlap on a continuum with respect to our control over them. So, the substantial distinction theory fails (again). Further, I argued that Holroyd and Kelly's suggestion that an implicitly biased interviewer cannot discern how her implicit bias influences her evaluation is as much a problem for agents influenced by *agential* preferences (such as the fondness for a job applicant who is also one's friend) as it is for the implicitly biased panellist.

We've now seen two sorts of cases where substantial distinction accounts on the basis of control over action fail: it turns out that we have both indirect, intervention control, *and*, direct, intervention control over implicitly biased

actions and belief-guided actions. There's a third and final sort of control that we have over implicitly biased and belief-guided actions, this time, a *non*-deliberative strategy for taking direct, intervention control, which I will now present in the final part of this section.

(ii) Non-deliberative, direct, intervention control

Agents may employ strategies to control their implicitly biased responses without knowing anything about the research on implicit bias. I suggest that these examples are best understood as cases of non-deliberative, direct, intervention control, as I will argue shortly. I will also outline some recent findings which indicate the neurological basis of a capacity to intervene (without the need for introspective awareness or deliberation) on motor processes that are already in motion when changes in the environment mean that a current action is no-longer inline with an agent's active goals (Aron, 2011). This shows that non-deliberative, direct, intervention control, which operates below the level of introspective awareness, may also play a crucial role in many of our agential actions.

A number of studies show that the extent to which agents manifest implicit bias in their actions covaries with various agential factors. For instance, Devine *et al.* (2002) show that agents who are already motivated to refrain from prejudice because they think that doing so is inherently valuable (Devine *et al.* call this an 'internalised' motivation) exhibit less implicit bias than agents who profess to thinking that refraining from prejudice is valuable because they are worried about how they are perceived, and less still than those who are not motivated at all. In fact, Devine *et al.* suggest that "the more internalized or self-determined a goal or value is, the more successful people are at responding consistently with the goal or value," (2002: 836). Succinctly, people who value responding without prejudice for the very reason that doing so is valuable in itself manifest less implicit bias in their actions than those who are motivated for instrumental reasons, or not motivated at all. Agents do not need to know anything about implicit bias, or related research on control strategies, to already be responding without prejudice as above, they just need to be motivated by the notion that prejudice is an inherently bad thing, which is, arguably, an agential characteristic.

Other research shows that agents with genuine long-term commitments to non-prejudice also manifest less implicit bias in their actions than those without such commitments (Moskowitz et al.,1999). This finding relies on earlier research

in which it was shown that when participants are made to engage in behaviour which violates what they indicate on self-report measures to be a genuinely held long-term commitment, they try to alleviate the conflict felt by overcompensating in line with their commitment later on—performing what have been termed 'incompleteness behaviours' (Gollwitzer *et al.*, 1982). As measuring long-term commitments in this way requires that the commitments in question are *exercised* to bring about commitment-congruent behaviour, it provides a more direct record of the commitment at issue than measures which rely on participants' own reports about their commitments (Moskowitz *et al.,* 1999: 169). Speaking of long-term commitments to non-prejudice, Moskowitz *et al*. maintain that:

> If people commit themselves to such self-defining goals, they are expected to make use of available opportunities to express the goal and to hold on to it even in the face of hindrances, barriers, and difficulties. (1999, 169)

Moskowitz *et al.* hypothesised that agents who perform egalitarian incompleteness behaviours after being made to participate in a non-egalitarian task would also tend to manifest less implicit bias in action. In this case, the non-egalitarian task was a questionnaire on which participants were only able to give stereotypical answers about women, and the incompleteness behaviour was measured by their response to a subsequent questionnaire about women (on which they could give egalitarian answers). Those considered to be high in incompleteness behaviour were those who responded with significantly more egalitarian responses after being forced to give stereotypical answers. As hypothesised, Moskowitz and colleagues observed a correlation between those who performed incompleteness behaviours after engaging in a non-egalitarian task (so, those with long-term commitments to egalitarianism) and those who manifested less implicit bias on another test. So, holding long-term egalitarian commitments enable agents to control their implicit social responses accordingly. A later study from Moskowiz and Li (2011) reveals that egalitarian commitments may be 'triggered' in participants without their realising, which then inhibit the manifestation of implicit bias in later behaviour.[56]

[56] These commitments were triggered in participants by making them contemplate a past event in which they failed to be egalitarian towards an African American man.

Interestingly, Moskowitz *et al.* (1999) maintain that the processes which bring implicit responses in line with an agent's long-term commitments are not conscious, or effortful, and do not require the agent to consider whether they still believe that they should refrain from prejudice whenever the relevant social concepts are made salient. Moskowitz *et al.* propose that the long-term egalitarian commitment in question operates automatically to prevent the facilitation of stereotypic categories in the presence of the relevant social concepts (1999: 168). Accordingly, agents who have cultivated longstanding commitments to egalitarianism have done so as a result of *already* having responded to reasons to refrain from prejudice. Such agents, it turns out, do not need to consciously consult these reasons each time they find themselves in a situation where they could act prejudicially or fairly, in order for the commitment to engender reasons-congruent behaviour. In this sense, the control implicated in these studies is non-deliberative in kind.

These results put pressure on Levy's (2014a: 53) claim if attitudes are activated in novel circumstances they must be conscious in order to be processed consistently with the agent's consciously held values.[57] On the contrary, these results show that attitudes may be activated and processed nonconsciously in novel circumstances to bring about behaviour that *is* consistent with the agent's consciously held values. Whilst suppressing bias for the above agents is neither the outcome of a deliberative process, nor available to introspection, because it correlates with the agent's internalised or genuine, long-term commitments, I suggest that bias suppression is best modelled as an agential action, which puts further pressure on the SD claim that implicitly biased actions are not agential on the basis of our (apparent) lack of control over them.

Holroyd and Kelly (2016) maintain that the utilisation of long-term commitments to calibrate responses to situational features in line with the agent's values is rightly considered as an agential capacity, even when such responses are not guided by attention. Speaking of Moskowitz and Li (2011) they suggest:

---

[57] Levy says "Activating concepts nonconsciously has effects on subjects' attitudes, but these effects are associative and not logical. All of this appears to be evidence of an absence of the capacity to integrate the content of representations; whereas nonconscious processing of contents may cause the activation of semantically related content, only when the processing is conscious is the activation logically coherent" (2014a: 53). I summarised these sentiments in claims NL2 and NL3.

> The agent's values and goals themselves, then, can play a role qua mechanisms that influence and calibrate the subsystems that run without reflective or direct control. This is a case of one element of a person's psychological economy influencing another. The agent's values 'keep in check' the operation of implicit bias, such that pursuing certain values is one way of exercising ecological control even when one is not actively monitoring one's actions with respect to whether they promote (or depart from) those values. Crucially, this can be so without the agent expressly intending, at any point, to put in place mechanisms for this purpose. (Holroyd and Kelly, 2016: 123)

Holroyd and Kelly (2016) claim that this kind of control is an example of *indirect* intervention control. But I am not sure that this is the correct model. Instead, I think this is an example of (non-deliberative) *direct,* intervention control. Indirect control, as we defined it above, is when an agent ψ-s in order to bring about that she ϕ-*s*. That is, the act which the agent themselves performs is *distinct* from that which she seeks to bring about. But when an egalitarian goal inhibits the activation of a stereotype, I am not sure which act is supposed to be the agent's ψ-ing and which her ϕ-*ing*: it seems to me that we have just one act: the inhibition of the stereotype. In §5.2.2, I gave examples of agents (Muhammed, Laura, Naveen, and those experiencing flow) who are not consciously aware that an attitude guides behaviour, and nor do they intend for this to be the case. I argued that, nonetheless, these attitudes are rightly identified as agential. I think that we have a similar situation here. If a persistent egalitarian motivation (which seems rightly identified as an agential state) guides or suppresses action, then, even if the agent in question is not aware that this is the case, and exerts no effort over its being so, it still strikes me that this is a case of *direct* control. There is just one act, the suppression of a stereotype and, because this is the result of an agential attitude, I think that this is rightly identified as something that agent herself does. Thus, I think the findings of Moskowitz *et al.* (1999), Devine *et al.* (2002) and Moskowitz and Li (2011) demonstrate that agents who are genuinely committed to egalitarianism may directly and non-deliberatively intervene on the manifestation of implicit bias in action. That said, I agree with Holroyd and Kelly's analysis of the control present in the Moskowitz and Li (2011), and similar experiments, insofar as they suggest that even though the agent doesn't introspectively micromanage every detail of how their values and commitments

are expressed in their behaviour, the agent is fundamentally implicated at the heart of this sort of control—it's no *accident* that egalitarian agents can effortlessly inhibit implicitly biased responses.

Moreover, as well as non-deliberative, direct, intervention control over implicitly biased actions, we may well have this kind of control over a variety of other agential actions—which is problematic for the SD theorist. Recent neuroscientific findings reveal a neurological basis for a form of non-deliberative, direct, intervention control over the motor behaviours that constitute many of our everyday actions. In a recent study, Aron (2011) observes that people are capable of 'spur of the moment' inhibitions of behaviour in response to changes in their environment. According to Aron, the speed at which study participants adjust their behaviour suggests that they are proactively inhibiting a process before it begins. Although participants do not deliberately guide the inhibitory response, and it occurs without effort, it is activated in a manner that is sensitive to participants' current goals—behaviour is inhibited when there is a sudden change in environmental stimuli, such that, had the behaviour in question gone ahead, it would contradict the participants' current goals. Because of this, Aron maintains that this inhibitory process is generated automatically in accordance with the agent's current goals (like it would seem to be in the Moskowitz et al. (1999) results, as above). Aron calls this "proactive inhibitory control" (2011).

Results in the 2011 paper are confined to laboratory tasks, but Aron suggests that a 'real-world' application of the proactive inhibitory control mechanism could be "preventing oneself from stepping into the street when the light changes color" (2011: 61). Presumably, the active goal here is something like 'crossing the street safely'. The idea is that if this goal is active, and a red light is detected, then the motor processes which generate walking will be automatically terminated (non-effortfully, on the part of the subject). Another instance in which the proactive inhibitory control mechanism may operate in the real world is in "sports requiring fast action control, such as stopping and switching movements in response to changing environmental signals" (2011: 61).

Such results would seem to paint a picture of agency where actions can manifest an agent's goals, values and commitments, without necessarily requiring deliberation, introspective awareness or effortful control. If that is right, then the point cuts both ways: if we wish to maintain that participants in the Moskowitz *et al.* (1999), Devine *et al.* (2002) and Moskowitz and Li (2011) experiments, as

well as agents utilising Aron's "proactive inhibitory control" (2011) act agentially, then it is difficult to argue that some agents who fail to suppress their implicit biases are not at all implicated as agents when they do: at the very least, they are failing to do something that, to judge from what other agents manage, they could and should do.

My arguments in this section have shown that, far from it being the case that we have no control over the manifestation of implicit bias in our actions, agents have three distinct effective strategies for controlling implicitly biased actions: (i) indirect, intervention control; (ii) deliberative, direct, intervention control; and (iii) non-deliberative, direct, intervention control. Moreover, I have argued that these strategies are by no means exclusive to the control of behaviour guided by implicit attitudes; rather, they are necessary for the control of at least some everyday agential actions, and, besides, these kinds of control are both regularly employed by, and even *expected* of many agents. We can conclude then that the SD theorist's claim that there is a kind of control that we have over all of our belief-guided actions which we do not have over our implicitly biased actions is false.

Summary

In the foregoing, I have argued that there is no substantial distinction between (i) the acquisition of implicit biases, and the actions that they influence; and (ii) the acquisition of agential attitudes, such as beliefs, and the actions that they guide, on the basis of the kind of control that we exert over each. In particular, I argued that there is no substantial distinction between the control that we exert over the acquisition of agential attitudes, such as beliefs, and that which we exert over the acquisition of implicit biases. I showed that if one thinks that we exert indirect voluntary control over belief acquisition and update, then, following a number of empirical findings, implicit bias acquisition and update can also sometimes be indirectly voluntary. I then showed that if one is committed to direct doxastic control (such as in Hieronymi's account of 'answerability', 2008), then one is also committed to direct control of at least some implicit biases. I then argued that there is no substantial distinction between the control that we exert over our agential actions and that which we exert over our implicitly biased actions. I demonstrated that three distinct strategies are effective for controlling implicitly

biased actions: indirect, intervention control; deliberative, direct, intervention control; and non-deliberative, direct, intervention control. Here, I argued that these kinds of control are also the *only* strategies available to us for controlling at least some of our everyday agential actions.

This result undercuts SDR arguments which proceed on the assumption that (a) we do not exert any control over implicit biases, and their influence on action, or that (b) we do not have the kind of control of our implicit biases, and their influence on action, that renders such attitudes and actions as agential. I showed (b) to be false, and the falsity of (b) includes the falsity of (a). In light of this, consider the following claims, summarised from the argument of SDR theorist Saul (2013):

> **JS5:** It is a necessary condition for moral responsibility for having a mental state *m*/for action influenced by a mental state *m* that the agent is able to control the acquisition of *m.*
>
> **JS6:** It is a necessary condition for moral responsibility for action influenced by mental state *m*, that when an agent becomes inferentially aware that she has *m*, she is instantly able to control the influence of *m* on action.

We are yet to determine the precise account of control that Saul has in mind in the above. However, it doesn't matter: Given that (i) we have some kinds of control over the acquisition of our implicitly biases, as well as over our implicitly biased actions, and (ii) these kinds of control are also the *only* strategies available to us for controlling at least some of our beliefs and our everyday agential actions, then neither JS5 nor JS6 rule out moral responsibility for implicit bias and implicitly biased actions—unless they *also* rule out moral responsibility for a great many other seemingly agential actions.

Now consider the following from Levy (2014a):

> **NL1:** …only when we are conscious of the facts that give our actions their moral significance are those actions expressive of our identities as practical agents and do we possess the kind of control that is plausibly required for moral responsibility, (2014a: 1).

Levy (2014a) argues that consciousness$_{PA}$ (his particular account of consciousness as personal availability) is necessary for the kind of control that is required for moral responsibility. As I argued at the end of Chapter 3, consciousness$_{PA}$ of the morally relevant facts is an overly demanding condition for agents to meet in order to be morally responsible, and a condition according to which a lot of intuitively responsible agents (such as the *explicitly* prejudiced) will turn out not to be responsible. I think that this is especially true for the agents that I discussed in the latter subsection of §5.3.2, whose motivations to refrain from prejudice were activated, enabling them to exert a non-effortful, non-deliberative form of intervention control over their implicitly biased actions, without any awareness that they do so (Moskowitz *et al.*, 1999; see also Devine *et al.*, 2002; and Moskowitz and Li, 2011). Intuitively, because an agential state (motivation) guides the inhibition of stereotypical responses in line with agents' commitments, these agents act agentially, even though they are not conscious$_{PA}$ of the moral significance of their action. If agents whose motivations guide their actions such that they refrain from prejudice are appropriate subjects of moral praise, it is not clear to me why the agents in the above experiments should be ruled out as praiseworthy (and those who fail to be motivated to refrain from prejudice, as blameworthy). I therefore conclude that we have grounds for rejecting NL1. Accordingly, if agents are not morally responsible for their implicitly biased actions, then it will not be because of their lack of control.

By this point in the dialectic, however, we have ruled out all of the supposedly distinguishing features that were introduced in Chapter 2 as able to uphold a substantial distinction between implicit biases, and implicitly biased actions; and agential attitudes and actions. This sets the stage for a more positive explication of the continuum thesis, and of the sense in which implicit biases, and implicitly biased actions, are both agential and morally evaluable. This will be the topic of the sixth and final chapter.

CHAPTER 6: AGENCY AND RESPONSIBILITY ON THE CONTINUUM THESIS

In the foregoing chapters, I argued that there is no substantial distinction (SD) on the basis of awareness, structure and processing, or control, between (i) implicit biases and implicitly biased actions; and (ii) beliefs and belief-guided actions. In particular, I showed that agents do have some awareness of the influence of their implicit biases on action, and further, that agents may lack awareness of their attitudes and their influence on action, and yet those attitudes and actions may still be agential. I then demonstrated that at least some implicit biases encode propositional information and, additionally, that beliefs do not always update in light of evidence. Finally, I argued that there is a number of strategies available for controlling implicit biases and related actions which are also the only strategies by which we control beliefs, and at least some agential actions.

In this final chapter, I present a more positive account of the nature of the attitudes that, following the arguments in the previous three chapters, we end up with. I defend the notion that agential attitudes and actions lie on a continuum in accordance with the level of awareness and control that we have over them, a continuum which is also populated by implicit attitudes, and the actions that they guide (§6.1). There is a large enough overlap on this continuum between the former and the latter that we should think that at least some of the latter are properly identified as agential. At this section of overlap on the continuum, if it is appropriate to hold agents as morally responsible for their agential actions (and I show that it is), then it is appropriate to hold agents as morally responsible for their implicitly biased actions, with which these agential actions share features. The continuum view also enables us to account for at least some of the considerations that motivate the SD view in the first place, without committing to the problems that it generates, as I will shortly argue.

One way that the SD theorist could try to respond to the case that I have been building against them is to argue that all of my counter-examples to the substantial distinction theory (counter-examples where apparently agential attitudes overlap with implicit attitudes along the continuum) are *not* agential, after all. To do this is to effectively raise the bar of agency higher up the attitude continuum so that the last implicit attitude (and any associated actions) fall

outside the domain of the agential. Call this the 'bar-raising' response. In §6.2, I argue that the bar-raising response cannot make sense of a number of occurrences which surround my counter-examples to the SD theory from the previous three chapters. Specifically, (i) it cannot make sense of many or our practices of praise or blame; and (ii) it cannot make sense of the notion that we learn something new about *ourselves* when we discover how our implicit preferences manifest (as argued by Smith, manuscript). The 'bar-raising' response, therefore, forces us to adopt an intolerably deficient account of agency, in which a significant set of human activities and flourishing turn out to be non-agential. This constitutes sufficient grounds for rejecting the bar-raising response, and consequently, the continuum thesis of implicit bias offers a superior account of the phenomena at hand to the substantial distinction theory.

6.1. The Continuum Theory of Agency and Moral Responsibility

In light of the failure of the substantial distinction account, what are we to say about the nature of implicit biases? I propose that the continuum thesis—on which implicit biases and agential attitudes such as beliefs (and the actions guided by each) do not have a fundamentally different nature—naturally accommodates the data presented in the last three chapters. Further, the continuum thesis is also able to account for a notion that I think motivated the SD argument in the first place: the idea that there is something like a 'gold standard' of agential attitude and action. However, the continuum thesis can do this without also jettisoning a range of other attitudes and actions which do not meet this standard, but which, intuitively, seem to be agential nonetheless. It is therefore a preferable account to the SD theory, as I demonstrate in the following.

6.1.1 Implicit biases and beliefs are not substantially distinguishable

During the course of this thesis, I have established that a considerable set of implicit biases and beliefs (and the actions guided by each) in fact share the very properties which other philosophers have tried to utilise to distinguish them as fundamentally different in kind. As a result, there is no principled way to draw a distinction between all implicit biases and all beliefs, and so what I have been calling the substantial distinction account fails. As such, it does not make sense to

maintain that implicit biases are a fundamentally different kind of attitude to beliefs.

This is not to have shown that implicit biases and beliefs are *identical.* Instead, it is consistent with what I have shown in the foregoing chapters that some beliefs may, for instance, figure frequently in an agent's introspective reasoning and planning; whilst some implicit biases may be incredibly difficult to observe, or to introspect on, without considerable deliberate self-reflection. But, what I *have* shown is that, in the middle of these two extremes, there are a number of beliefs and implicit biases of which agents have not yet become introspectively aware, but which share enough of their properties with attitudes which we do think of as agential that we do not have sufficient grounds for rejecting the former from the set of agential attitudes.

Further, it is consistent with what I have presented here that there are some belief-guided actions which proceed from episodes of careful conscious deliberation, and over which we exercise immediate direct, initiation control, whilst many implicitly biased actions occur in the absence of deliberation about performing such an action, and it may take considerable effort before they are amenable to any kind of control. Nonetheless, in the middle of each of these two extremes, are a number of belief-guided actions, and implicitly biased actions, which are not the result of effortful, introspective processes but which share enough of their properties with actions which we do think of as agential that we do not have sufficient grounds for rejecting the former from the set of agential actions.

This area of overlap between implicit biases and beliefs (and their associated actions) is considerable: such an area was found for every potentially distinguishing feature proposed by prominent SD theorists, such that the feature in question *failed* to uphold a substantial distinction between implicit biases and beliefs (and their associated actions). As we have seen, as regards awareness, agents may have introspective awareness of their implicit biases (such as Borgoni's case of Emilia from Chapter 3), or, if one rejects Borgoni's account of introspective awareness, then agents will *fail* to have introspective awareness of at least some of their agential attitudes (those which constitute their everyday observable class preferences). Further, agents may have observational awareness of many implicit biases (and agential preferences) in virtue of reflecting on how

these manifest in behaviour. So awareness will not deliver the desired distinction for the SD theorist.

We also saw that at least some implicit biases are structured propositionally, and may figure in inferential transitions, just as beliefs do (Chapter 4). This being so gives us reason to doubt the predictions of the theoretical model (dual process theory) on which implicit biases were thought to be associative in the first place. Further, it is not clear what inferential sensitivity has to do with agency and moral responsibility. Recall that Liz, the agent with a recalcitrant (explicit) racist prejudice from the end of Chapter 4, seems to be a target for moral condemnation precisely because her attitude has a very low degree of inferential sensitivity, and she fails frequently to update her racial prejudices in light of counter-evidence. (For this reason, I don't think that the level of inferential sensitivity of an attitude is reliable criteria by which to judge whether it is agential or not.)

Furthermore, we saw that we can have the same kind of control over the acquisition and maintenance of implicit biases (whether that is indirect voluntary control, or direct control in the form of answerability) as we do over the acquisition and maintenance of beliefs. Finally, there are three kinds of control strategies which are effective over implicitly biased actions, which are also the *only* strategies that we may use to control a number of everyday agential actions: indirect, intervention control; deliberate, direct, intervention control; and non-deliberate, direct, intervention control. So, even after testing multiple notions, control will not deliver the desired distinction for the SD theorist.

One might think that there may be other potential SD arguments out there that I have not assessed in this thesis, and point out that the claim that there are *no* distinguishing criteria between all implicit biases and beliefs (and their associated actions) hasn't quite yet been established. If so, there might be hope for the SD theory yet. But I am not hopeful that the SD theory can be saved with any further possible distinctions. Part of the thrust of my argument has been to show that implicit biases are, fundamentally, not that unlike beliefs. But, perhaps even more importantly, over the course of this thesis we have seen that *beliefs* themselves are a somewhat messy class of attitudes, and far from all of them measure up to the thoroughly rationalistic picture that philosophers such as Gendler and Levy espouse. So, I think that some of the same cases of as yet unobserved, inferentially-insensitive beliefs, as well as the not directly controllable actions that

they guide, will prove to be problematic for any further potentially distinguishing criteria which I have not directly assessed in the last three chapters.

So, for any possible distinguishing criteria, we are left with a substantial area of overlap between implicit biases and agential attitudes. Either at least some putatively agential attitudes in fact *fail* to have the criteria which would distinguish them as agential, or at least some implicit attitudes turn out to *have* criteria which render them agential after all. This marks the failure of the substantial distinction account. In the next section, I outline an account that can make sense of the picture that we are left with: the continuum thesis.

6.1.2. The continuum thesis

The proposal, then, is this: Agency is a property that comes in degrees, and there is a continuum along which attitudes and actions are ordered, from the least agential, to the most agential. That agency comes in degrees is not a particularly radical idea, and has been defended previously by Nahimas, (2006). Nahimas suggests that the idea that agency comes in degrees fits naturally with some of our established concepts and practices. For instance, it makes sense of our notion of children as developing agents, and our practice of responding to others in accordance with varying degrees of praise and blame with respect to the (moral) severity of an action.

On the continuum that I propose, there are two main dimensions; those of awareness and control (as I indicated above, I don't think that inferential sensitivity is a particularly informative heuristic to agency, because highly recalcitrant explicit prejudice will end up being non-agential, but, as I demonstrated in Chapter 4, the inferential sensitivity of an explicit racial prejudice, for instance, seems to have little to do with whether the attitude is agential, and whether we blame the agent in question). Attitudes are ordered along these two dimensions such that those closer to the higher end of the awareness scale, or the control scale, are to be understood as proportionally more agential than those toward the lower end of the awareness or control scales. Many agential attitudes and actions will occupy roughly equivalent co-ordinates on the awareness dimension as they do on the control dimension, but this is not the case for *all* examples: for instance, cases of 'flow' will be much higher on the control dimension than they are on the awareness dimension.

Some highly calculated moral wrongdoings will lie at the higher end of the awareness and control dimensions, and some behaviours which are influenced automatically by attitudes of which it is very difficult to become aware without a great introspective or observational effort lie at the lower end. In the middle, however, there is a significant area of overlap between implicit biases (and the actions that they guide) and beliefs (and the actions that they guide): In this area we find the set of in principle observable (but currently unobserved) everyday class preferences (Chapter 3), as well as some implicit biases that are observable class preferences. We also find the everyday unrestriction cases and examples of flow (Chapter 5), as well as some implicitly biased actions. The extent to which these states and actions are agential, and the severity of our reactive attitudes when such states guide moral wrongdoings depend on the position that they occupy on the continuum.

In this region of overlap between what has been called the 'implicit' and what has been called the 'explicit', if it is appropriate to hold agents as morally responsible for their agential actions, then it is appropriate to hold agents as morally responsible for their implicitly biased actions, with which these agential actions share features. I argued that there are plenty of cases in which agents both can, and are in fact *expected*, to, for instance, (i) notice that their preferences automatically guide their actions in a fashion inconsistent with their aims; and (ii) redirect their future behaviour in line with a non-preferential course of action (in Chapter 5). For example, when the student council chair's left-wing preference manifests in the selection of questions from known left-wingers at a higher rate than questions from known right-wingers during a council meeting, it is possible for (and expected of) her to notice this, and to (directly) intervene and actively look for questions from known right-wingers.

Similar expectations would then appear to apply to the implicitly biased seminar chair, whose case shares characteristics with respect to awareness and control with that of the student council chair (Chapter 5): It is also quite possible for the seminar chair to notice his predominant selection of questions from men, and to directly intervene on the manifestation of this preference, by actively looking for questions from women and people of colour. Individuals in these cases are the proper subjects of our reactive moral attitudes, insofar as they are in positions of power where they are expected to act *fairly.* When they fail to do so, even if this failure is neither voluntary, nor the product of deliberate, direct,

initiation control, (and, rather, is explained by the automatic manifestation of a preference) it seems entirely appropriate to ask these agents to account for their actions, as well as to respond to these agents with attributions of blame. That is, although these actions may be the product of automatic processes, our reactive attitudes address the *agent*, we ask the *agent* to account for these actions. So, whilst we might not think that actions in the overlap region, should they violate some moral norm, demand the same severity of blame as highly calculated actions, there is still room for holding agents morally responsible for acting as they do, at least to some extent.

At this point, let me say a few words about Washington and Kelly's (2016) account of moral responsibility for implicit bias, in order for me to acknowledge some similarities, and to also demonstrate how my suggestions above differ to the account that they put forward. Washington and Kelly argue that what people ought to be aware of with respect to implicit bias (and to therefore employ control strategies against) is indexed to the kind of role that they play in society. Specifically, they are interested in the notion of *inferential* awareness (as Holroyd 2015 uses the term): that is, awareness of the empirical findings on implicit bias. Specifically, Washington and Kelly argue that people in what they term 'gate keeper' positions, such as those involved in hiring committees, education or social work, for instance—positions which involve the fair distribution of social resources, enabling other people to self-determine—ought to have inferential awareness of the relevant empirical findings on implicit bias, and ought to employ suitable mitigation strategies in their actions. To the extent that they fail to do this, they may be held morally responsible.

Holroyd (2015) argues that Washington and Kelly's proposal is implausible, because:

> Many people make decisions about…who to grant a loan to, where to live, who to stop and search, who to give a lift to, what news stories to report (and how), who to write prescriptions for, who to sit by on a train, how to evaluate co-workers, who to smile at, what grades to assign or references to write, who to cross the road to avoid, who to believe, who to befriend…and so on (2015: 517).

Accordingly, almost everyone will turn out to play a 'gate-keeper' role in a variety of social interactions, but it is unreasonable to suggest that, therefore,

everyone ought to have inferential awareness of a particular set of findings in cognitive science as regards implicit bias (Holroyd, 2015: 517).

I agree with Holroyd's (2015) contention here, but insofar as my proposal differs from Washington and Kelly's (2016), this does not affect my argument. My argument above requires that the student council chair and the seminar leader are able to have *observational* awareness of the manifestation of their preferences. I think that it is quite plausible that they can have this sort of awareness, as I argued in Chapter 5. (Further, if you are convinced by Borgoni's (2015) 'ordinary' account, then it is possible that they have *introspective* awareness of their preferences.) As such, it doesn't matter if it is implausible to expect the seminar leader, and others like him, to know about the relevant empirical findings of implicit bias, because it is relatively easy for him to *observe* that he is only selecting questions from white men, particularly when the brief of his role is to select questions fairly. So, if it is plausible to expect the politically biased student council chair to observe and correct the bias in her own behaviour, and to hold her to account if she fails to (and, as I have argued, I think that it is) then it is equally plausible to hold the seminar chair to account for his biased question selection.

So, the continuum account does offer a framework on which, at least sometimes, agents will be morally responsible for their implicitly biased actions: They will be morally responsible when it is appropriate to hold other agents whose actions share fundamental features with the implicitly biased actions in question as morally responsible for their actions. Both sets of agents will be morally responsible to a similar degree, and appropriate subjects of the same level of moral condemnation.

There is a further benefit for continuum theorists. As well as being able to account for the examples that I presented in Chapters 3-5, examples that the substantial distinction view cannot account for, that beliefs and implicit biases lie on a continuum also enables us to acknowledge at least some of the considerations that I think motivated the substantial distinction view in the first place: One may be attracted to the idea that there is (a) a 'gold-standard' of agential attitude, in which attitudes are acquired and updated in accordance with evidence, and under the guidance of occurrent, introspective awareness; as well as (b) a 'gold-standard' of agential action, in which agents deliberate about how to act, whilst being occurrently, introspectively aware of the relevant set of attitudes, and the guiding role that they play in directly initiating an action. One might then also

think that when these sorts of actions result in the violation of moral norms, then particularly severe reactive attitudes, which emphasise the deliberateness of the norm-violating action, are appropriate from others. However, a commitment to the notion that agential actions which violate moral norms in this way are rightly to be met with severe moral condemnation, is fully consistent with the acknowledgment that both agency and moral responsibility may be a matter of degree, determined by the position of the relevant attitudes and actions on the continuum. These kinds of actions will appear high up the continuum, and so may be met with severe moral condemnation. Continuum theorists can acknowledge, and account for this idea, without also being committed to a substantial distinction which has proved so problematic to maintain in a principled way.

So, on the continuum view, the term 'implicit' does not pick out any particular characteristics of an attitude, it merely indicates the region in which it is likely to lie on the continuum. One end of the ordered of the set of things which we have been calling 'implicit' is independent of one of the ends of the ordered set of things that we have been calling 'explicit'. However, these sets overlap significantly at their other ends. As such, our best model of the phenomena is that, as categories of attitude, the implicit and the explicit are not discontinuous from one another, and indeed, have a considerable intersection. Further, at least sometimes, agents will turn out to be morally responsible for implicitly biased actions: when those actions share fundamental characteristics with other attitudes for which we already have a precedent for holding agents as morally responsible.

I acknowledge that SD theorists might still not be convinced, and may have a last line of defence. I consider, and respond to this in the final section of this chapter.

6.2. The Bar-Raising Response

In light of the foregoing, it remains a dialectical possibility that SD theorists will respond by insisting that the argument that I have presented in this thesis sets the bar of agency too low. The suggestion might be that none of the attitudes and actions that I present which do not meet the 'gold-standard' of agency really count as genuinely agential. If that is so, then my argument that there is no substantial distinction between implicit biases and agential attitudes (and the actions guided by each) does not have the required consequence for it will then turn out that *no* implicit biases, and implicitly biased actions, count as agential after all.

However, this response demands that we reject an intolerably large variety of human attitudes and actions from the category of the agential. Let us now look at all of the attitudes and actions that the bar-raising response requires us to reject as agential in summation:

(i) agents who have as yet unobserved, but nonetheless observable class preferences, which, as I argued in Chapter 3, characterise not just some of our implicit biases, but a great many of our everyday aesthetic and prudential preferences as well;
(ii) agents with recalcitrant beliefs which do not update in light of evidence, which characterises a great many *explicit* prejudicial beliefs, as I demonstrated in Chapter 4;
(iii) agents who put in place implementation intentions to indirectly control everyday actions as we saw in Chapter 5, for example to enhance their emotional expression (such as actors) or to steel themselves in traumatic situations (such as accident and emergency doctors);
(iv) agents who recognise and redirect the automatic manifestation of personal preferences in their behaviour in situations in which they ought to be acting non-preferentially, such as the student council chair, and the TV planner from §5.2.2;
(v) sports players and musicians performing highly skilled, novel action sequences without deliberative or introspective guidance during episodes of 'flow', such as in §5.2.2.

On the bar-raising response, these agents are neither acting agentially, nor are they praiseworthy or blameworthy (as appropriate) for their actions. I contend that this is an intolerably large set of human behaviour to jettison from the class of agential attitudes and actions. It contains behaviour that regularly characterises everyday actions, and further, behaviour that we might think represents some of the pinnacles of human flourishing and achievement (such as skilled musicianship and sporting accomplishments).

But we have a further problem should we jettison the above examples from the set of agential actions: doing so renders much agential practice that often accompanies many of the above actions as inappropriate or meaningless. For

instance, on the bar-raising response, it would be inappropriate to praise the skilled trumpet player for her incredible, complex, virtuosic ten-minute solo. And yet, we regularly do praise musicians for such feats. Further, our praise seems to be almost entirely independent of whether or not the musician in question was in a 'flow' state (without occurrent, introspective awareness of the non-deliberate processes which guide their playing) or not. If they were in a flow state, it is not clear how, if at all, this would impact our appraisal of their playing. Nor does the bar-raising response make sense of at least some of our blame practices: according to the bar-raising response, it is inappropriate to hold the student council chair as both morally and politically accountable for favouring her allies, and inappropriate to blame the climate skeptic TV planner for programming significantly more airtime for climate skeptics than climate scientists, (Chapter 5). But, as I argued in the previous section, we *do* hold such agents to account, and so the bar-raising response does not make sense of our practices here. The bar-raising response fairs no better in the moral realm: According to this response, it is inappropriate to blame Liz, the recalcitrant racist from Chapter 4, for failing to update her attitudes in light of the evidence that she takes herself to have for doing so. And yet, as I suggested in Chapter 4, and above, the degree of evidence sensitivity of an attitude is not obviously related to moral responsibility in the case of agents with a recalcitrant explicit prejudice.

Smith (manuscript) points out a further problem for those who think that attitudes and actions like those listed at the start of this section are non-agential. Smith maintains that we feel like we learn something new *about ourselves* when we notice a previously unconscious preference, and how it manifests in behaviour (manuscript). She suggests that "[s]uch discoveries are very different from coming to discover that we have cancer, or high blood pressure, or any other physical ailment or condition" (Smith: manuscript: 20). Whilst in these latter cases, we are making discoveries about the operation of non-voluntary, non-deliberative processes, they are a different *kind* of discovery to recognising non-voluntary, non-deliberative aspects of one's evaluations. Even though these evaluations may be non-voluntary, and non-deliberative, they still "embody our take on the world" as Hieronymi would say, (2008: 370) in a way that a physical ailment does not.

Indeed, agents who observe their (until that moment) unobserved class preferences, as well as agents who realise that an already observed preference

manifests automatically in contexts in which they did not intend it to, would seem to learn something new about what matters to them, and about what they act on as valuable. For instance, the student council chair learns something about herself when she recognises that she favours questions from known left-wingers during a council session—perhaps she learns something about the depth of her commitment to the left-wing of politics, her desire for her political allies to be heard, and how these desires can overshadow her commitment to fair a debate. Similarly, it would seem that the seminar chair learns something important about himself when he realises that he predominantly selects questions from white men—that he doesn't value the contributions of women and people of colour as much as he does from white men. It might be as uncomfortable for him to learn this as it is for Clare (from Chapter 3) to discover that that genre of music that she likes from her recent festival experience is the very same genre that her friends abhor, but, nonetheless, both Clare, and the seminar leader, discover aspects of their own evaluative agency when they observe the manifestation of their preferences. As Smith argues, if we really are mistaken in attributing the above attitudes and actions to ourselves as agents, "then these impressions of self-discovery and increased self-knowledge must be illusory" (manuscript: 20) and this is an outcome that is hard to reconcile with the phenomena at hand.

The bar-raising response also seems inadequate to capture our regular, everyday moral practice, where, as Smith points out (manuscript), our reactive attitudes and attributions of blame are not restricted to actions which are the products of deliberate, direct, initiation control. Nonetheless, such reactions are comprehensible. She maintains that:

> …we sometimes ask people to justify their failures to notice, to remember, or to take into consideration certain factors at the time they acted, when such failures were clearly neither "conscious" nor the result of prior conscious activities… But I think it is interesting that we generally do not regard such requests for justification as obviously infelicitous or bizarre when they are directed to us in response to such unconscious failures. "Didn't you realize how inappropriate that joke was in that context?," "How could you have missed the obvious warning signs in her behavior"?," "What could have been so important this afternoon that you forgot to pick up our daughter from school?" These are types of justificatory request we regularly make and

> receive, which suggests that there is nothing *conceptually* untoward about them. (Smith, manuscript: 18-9)

In light of this, we should reject the 'bar-raising' response. The theoretically preferable option, that is, the option that best explains the phenomena from our moral and agential practices as regards the belief-implicit bias overlap of the continuum, is to accept that these attitudes and actions are agential after all, and that, at least sometimes, it is appropriate to praise or blame the agents in question. If that is the case, then, as I argued in Chapters 3-5, at least some implicit biases, and implicitly biased actions, are agential—as agential as the attitudes and actions in examples listed at the start of this section, and it is therefore appropriate to hold the agents in question as morally responsible for them. This is not to be committed to the notion that it is inappropriate for the moral community to meet a calculated and deliberate violation of moral norms with particularly severe sanctions in accordance with the moral severity of the action in question. As I outlined in the previous section, the continuum account is *consistent* with the notion that motivates the SD theorist, that when agents deliberately and calculatedly violate (moral) norms, they are to be met with a particularly severe level of riposte. However, we need not commit to an intolerably restrictive account of agency—the account of agency of the bar-raising response—to accommodate this intuition.

On this rejoinder to the bar-raising response, agents who perform involuntary, non-deliberate actions (including some implicitly biased agents) may also, at least sometimes, act agentially, and be the proper subjects of moral praise and blame. So, we ought to reject the bar-raising response to the continuum thesis, and maintain that the continuum thesis is the best interpretation of the phenomena that have been the focus of this discussion.

Conclusion

A significant body of empirical evidence reveals that people often act as if they have negatively evaluated members of a particular social group, even though they seem to be unaware that this is the case, and do not intend to exhibit such disfavouring treatment (Chapter 1). Psychologists and philosophers alike have treated these so called 'implicit' biases, and the actions that they influence, as distinct in kind from our beliefs and belief-guided actions, and a subset of

philosophers have argued that implicit biases, and implicitly biased actions, are therefore not agential (Chapter 2).

In this thesis, I demonstrated that there is no principled way to maintain a substantial distinction between implicit biases and implicitly biased actions, and agential attitudes and actions, such that all of the former fall on one side of the distinction, whilst all of the latter fall on the other. I looked at the possible distinguishing features put forward by various substantial distinction theorists: those on the basis of (i) awareness; (ii) structure and processing; and (iii) control, and found that there is no single characteristic that all agential attitudes and actions have, that all implicit biases, and implicitly biased actions lack. I argued that we often have observational awareness (and, perhaps, introspective awareness) of those implicit biases which constitute observable class preferences (Chapter 3); that at least some implicit biases are sensitive to propositional information, and additionally, that beliefs do not always update in light of evidence (Chapter 4); and that there are a number of strategies available for controlling implicit biases and implicitly biased actions which are also the *only* strategies by which we control beliefs, and at least some agential actions (Chapter 5).

In showing the substantial distinction account to be inadequate, I defended a continuum thesis, on which implicit biases and implicitly biased actions, and agential attitudes and actions lie on a continuum. There is a significantly large overlap between these two categories on this continuum that we should think that at least some implicit biases and implicitly biased actions are properly identified as agential. At this point on the continuum, if it is appropriate to hold agents as morally responsible for their agential actions, then it is appropriate to hold agents as morally responsible for their implicitly biased actions, with which these agential actions share features. I argued that, to the extent that we *do* hold agents to account, and praise or blame them for agential actions which lie in the overlap zone of the continuum, it is therefore appropriate to hold agents as morally responsible for the implicitly biased actions which also populate that section of the continuum (Chapter 6). I then considered a possible dialectical move open to the SD theorist which I called the 'bar-raising' response, and argued that this response ought to be rejected because it commits us to an inadequate picture of agency, and cannot make sense of a number of the agential practices which accompany the discovery and manifestation of mid-continuum attitudes. As such,

the continuum thesis remains the account that is best able to accommodate both the findings on implicit bias and our moral practices regarding similar attitudes and actions.

On my account, implicit biases need not be viewed as central or defining features of our evaluative agency. However, it would be wrong to suggest that, therefore, they lie determinately outside the boundaries of agency, and that we, as agents, are wholly absolved from harbouring or acting upon them. This account encourages us to reflect upon the more recalcitrant aspects of our psychology, as well as to observe our behaviour to learn about the things that we value (and those that we do not value quite so much) in order to bring our attitudes and actions better in line with the motivations and values that we do take to be the defining aspects of ourselves.

AVENUES FOR FURTHER RESEARCH

The account that I have presented in this thesis generates a number of avenues for further research, as well as having implications for positive changes in practice as regards educating people about, and aiming to reduce, implicit bias.

In this thesis, I have defended a continuum view with respect to implicit bias. But cognitive science reveals that we harbour a great many attitudes, in addition to our implicit biases, which do not reflect the highly rationalistic picture of the mind that some philosophers have previously assumed, as well as revealing that we utilise these attitudes frequently in daily decision making and action. For example, self-deception, distorted memories, confabulatory explanations, and cognitive and informational biases more generally, are as widespread and ubiquitous as implicit social biases. Some have assumed that insofar as these cognitions are irrational, or ill-grounded, that they are not agential. Given the account that I have just presented, however, it is far from obvious that we are not acting agentially when these imperfect cognitions inform our behaviour. So, the continuum view may well have wider application, and be informative in at least some, if not all, of the above cases.

There are also implications for psychological theory. For example, in Chapter 4, I mentioned in footnote 34 that if implicit biases turn out to be propositionally structured, then this might have general implications for the dual process theory, because many dual process theorists maintain that implicit biases are paradigm associative states. I did not have room to explore this in the thesis,

but the rise of propositionalist interpretations of much of the dual process theorist's results (such as those of De Houwer, 2014 and Mandelbaum, forthcoming) opens up the possibility that the propositionalist model may well provide a better explanation of the non-social implicit attitude data than the dual process theory.

Finally, there is a more general practical significance to the outcome of this thesis, that has to do with communicating what implicit biases are to people who have never heard of them. The result that implicit biases aren't just quirks of our psychology that have little to do with us as agents, but instead, are not fundamentally different to the attitudes and values that we take ourselves to have, may have important motivational implications for how individuals and institutions tackle implicit bias. As we saw in the thesis, some methods for controlling implicit bias require more effort than others. However, if people see themselves as implicated in their implicit prejudices, then this may motivate them to take effortful steps to control the manifestation of implicit bias in action, as well as to aim to eliminate their implicit biases altogether. If institutions have to accept that their staff, and so themselves, are implicated in any implicit prejudices that arise within the workplace, then it might motivate them to investigate avenues for reducing implicit bias, as well as for instigating implicit bias training, if they haven't already, and to regularly review the effectiveness of these procedures, which ought to lead to a reduction in implicit bias, and greater awareness of the issue.

~

BIBLIOGRAPHY

Alvarez, M. 2009. "How Many Kinds of Reasons?" *Philosophical Explorations* 12 (2): 181–93.

Anderson, A. 1983. "Semantic and Social-Pragmatic Aspects of Meaning in Task-Oriented Dialogue." PhD thesis, University of Glasgow.

Anscombe, G. E. M. 1957. *Intention*. Cambridge, Massachusetts: Harvard University Press.

Armstrong, D. M. 1968/1994. "Introspection." In *Self-Knowledge*, edited by Q. Cassam. Oxford Readings in Philosophy. Oxford; New York: Oxford University Press.

Aron, A. R. 2011. "From Reactive to Proactive and Selective Control: Developing a Richer Model for Stopping Inappropriate Responses." *Biological Psychiatry* 69 (12): 55–68.

Aronson, E., and V. Cope. 1968. "My Enemy's Enemy Is My Friend." *Journal of Personality and Social Psychology* 8 (1): 8–12.

Arpaly, N. 2003. *Unprincipled Virtue: An Inquiry into Moral Agency*. Oxford: Oxford Univ. Press.

Banse, R., J. Seise, and N. Zerbes. 2001. "Implicit Attitudes towards Homosexuality: Reliability, Validity, and Controllability of the IAT." *Zeitschrift Für Experimentelle Psychologie: Organ Der Deutschen Gesellschaft Für Psychologie* 48 (2): 145–60.

Bargh, J. A., and E. Morsella. 2008. "The Unconscious Mind." *Perspectives on Psychological Science* 3 (1): 73–79.

Bargh, J. A. 1999. "The Cognitive Monster: The Case against the Controllability of Automatic Stereotype Effects." In *Dual-Process Theories in Social Psychology*, edited by S. Chaiken and Y. Trope. New York: Guilford Press.

Baumeister, R. F., and E. J. Masicampo. 2010. "Conscious Thought Is for Facilitating Social and Cultural Interactions: How Mental Simulations Serve the Animal-Culture Interface." *Psychological Review* 117 (3): 945–71.

Bennett, J. 1990. "Why Is Belief Involuntary?" *Analysis* 50 (2): 87–107.

Bertrand, M., and S. Mullainathan. 2004. "Are Emily and Greg More Employable than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination." *The American Economic Review* 94 (4): 991–1013.

Blair, I. V., J. E. Ma, and A. P. Lenton. 2001. "Imagining Stereotypes Away: The Moderation of Implicit Stereotypes through Mental Imagery." *Journal of Personality and Social Psychology* 81 (5): 828–41.

Borgoni, C. 2015. "On Knowing One's Own Resistant Beliefs." *Philosophical Explorations* 18 (2): 212–25.

Boyle, M. 2009. "Active Belief." *Canadian Journal of Philosophy* 39 (sup1): 119–47.

Broughton, R., R. Billings, R. Cartwright, D. Doucette, J. Edmeads, M. Edwardh, F. Ervin, B. Orchard, R. Hill, and G. Turrell. 1994. "Homicidal Somnambulism: A Case Report." *Sleep* 17 (3): 253–64.

Budden, A. E., T. Tregenza, L. W. Aarssen, J. Koricheva, R. Leimu, and C. J. Lortie. 2008. "Double-Blind Review Favours Increased Representation of Female Authors." *Trends in Ecology & Evolution* 23 (1): 4–6.

Chen, M., and J. A. Bargh. 1997. "Nonconscious Behavioral Confirmation Processes: The Self-Fulfilling Consequences of Automatic Stereotype Activation." *Journal of Experimental Social Psychology* 33 (5): 541–60.

Collins, A. M., and E. F. Loftus. 1975. “A Spreading-Activation Theory of Semantic Processing.” *Psychological Review* 82 (6): 407–28.
Csikszentmihalyi, M. 1990. *Flow: The Psychology of Optimal Experience*. Harper Perennial Modern Classics. New York: Harper and Row.
Dasgupta, N., and A. G. Greenwald. 2001. “On the Malleability of Automatic Attitudes: Combating Automatic Prejudice with Images of Admired and Disliked Individuals.” *Journal of Personality and Social Psychology* 81 (5): 800–814.
Dasgupta, N., and L. M. Rivera. 2006. “From Automatic Antigay Prejudice to Behavior: The Moderating Role of Conscious Beliefs about Gender and Behavioral Control.” *Journal of Personality and Social Psychology* 91 (2): 268–80.
Davidson, D. 1984. “First Person Authority.” In *Subjective, Intersubjective, Objective*. Oxford: Clarendon Press.
———. 1971. “Agency.” In *Essays on Actions and Events*, edited by D. Davidson. Oxford University Press.
Dawkins, R. 1976. *The Selfish Gene*. New York: Oxford University Press.
De Houwer, J. 2003. “The Extrinsic Affective Simon Task.” *Experimental Psychology* 50 (2): 77–85.
———. 2006. “Using the Implicit Association Test Does Not Rule out an Impact of Conscious Propositional Knowledge on Evaluative Conditioning.” *Learning and Motivation* 37 (2): 176–87.
———. 2014. “A Propositional Model of Implicit Evaluation: Implicit Evaluation.” *Social and Personality Psychology Compass* 8 (7): 342–53.
Dennett, D. C. 1987. *The Intentional Stance*. Cambridge, Mass: MIT Press.
———. 1991. *Consciousness Explained*. Boston: Little, Brown and Co.
———. 1995. *Darwin’s Dangerous Idea: Evolution and the Meanings of Life*. New York: Simon & Schuster.
Descartes, R. 1694. *The Passions of the Soul*. Translated by S. Voss. Indianapolis: Hackett Pub. Co.
Deutsch, R., B. Gawronski, and F. Strack. 2006. “At the Boundaries of Automaticity: Negation as Reflective Operation.” *Journal of Personality and Social Psychology* 91 (3): 385–405.
Devine, P. G. 1989. “Stereotypes and Prejudice: Their Automatic and Controlled Components.” *Journal of Personality and Social Psychology* 56 (1): 5–18.
Devine, P. G., E. A. Plant, D. M. Amodio, E. Harmon-Jones, and S. L. Vance. 2002. “The Regulation of Explicit and Implicit Race Bias: The Role of Motivations to Respond without Prejudice.” *Journal of Personality and Social Psychology* 82 (5): 835–48.
Dovidio, J. F., N. Evans, and R. B. Tyler. 1986. “Racial Stereotypes: The Contents of Their Cognitive Representations.” *Journal of Experimental Social Psychology* 22 (1): 22–37.
Dovidio, J. F., and S. L. Gaertner. 2004. “Aversive Racism.” In *Advances in Experimental Social Psychology*, edited by J. M. Olson and M. P. Zanna, 36:1–52. Elsevier.
Dovidio, J. F., K. Kawakami, C. Johnson, B. Johnson, and A. Howard. 1997. “On the Nature of Prejudice: Automatic and Controlled Processes.” *Journal of Experimental Social Psychology* 33 (5): 510–40.
Fazio, R. H. 1990. “Multiple Processes by Which Attitudes Guide Behavior: The Mode Model as an Integrative Framework.” *Advances in Experimental Social Psychology*, 23: 75–109.

Feldman, R. 2001. “Voluntary Belief and Epistemic Evaluation.” In *Knowledge, Truth, and Duty*, edited by M. Steup. New York: Oxford University Press.
Fischer, J. M. 1999. “Recent Work on Moral Responsibility.” *Ethics* 110 (1): 93–139.
Fischer, J. M. and M. Ravizza. 1998. *Responsibility and Control: A Theory of Moral Responsibility*. Cambridge Studies in Philosophy and Law. Cambridge: Cambridge Univ. Press.
Frankfurt, H. G. 1971. “Freedom of the Will and the Concept of a Person.” *The Journal of Philosophy* 68 (1): 5.
Gaertner, S. L., and J. P. McLaughlin. 1983. “Racial Stereotypes: Associations and Ascriptions of Positive and Negative Characteristics.” *Social Psychology Quarterly* 46 (1): 23–30.
Gawronski, B., and G. V. Bodenhausen. 2006. “Associative and Propositional Processes in Evaluation: An Integrative Review of Implicit and Explicit Attitude Change.” *Psychological Bulletin* 132 (5): 692–731.
———. 2011. “The Associative-Propositional Evaluation Model: Theory, Evidence, and Open Questions.” *Advances in Experimental Social Psychology* 44: 59–127.
———. 2014. “Implicit and Explicit Evaluation: A Brief Review of the Associative-Propositional Evaluation Model: APE Model.” *Social and Personality Psychology Compass* 8 (8): 448–62.
Gawronski, B., and B. K. Payne, eds. 2010. *Handbook of Implicit Social Cognition: Measurement, Theory, and Applications*. New York: Guilford Press.
Gawronski, B., E. Walther, and H. Blank. 2005. “Cognitive Consistency and the Formation of Interpersonal Attitudes: Cognitive Balance Affects the Encoding of Social Information.” *Journal of Experimental Social Psychology* 41 (6): 618–26.
Gendler, T. S. 2008a. “Alief and Belief.” *The Journal of Philosophy* 105 (10): 634–63.
———. 2008b. “Alief in Action (and Reaction).” *Mind & Language* 23 (5): 552–85.
Gladwell, M. 2005. *Blink: The Power of Thinking without Thinking*. London: Penguin Books.
Gollwitzer, P. M., and P. Sheeran. 2006. “Implementation Intentions and Goal Achievement: A Meta-analysis of Effects and Processes.” In *Advances in Experimental Social Psychology*, 38:69–119.
Gollwitzer, P. M., R. A. Wicklund, and J. L. Hilton. 1982. “Admission of Failure and Symbolic Self-Completion: Extending Lewinian Theory.” *Journal of Personality and Social Psychology* 43 (2): 358–71.
Goodin, R. E. 1995. *Utilitarianism as a Public Philosophy*. Cambridge; New York: Cambridge University Press.
Green, A. R., D. R. Carney, D. J. Pallin, L. H. Ngo, K. L. Raymond, L. I. Iezzoni, and M. R. Banaji. 2007. “Implicit Bias among Physicians and Its Prediction of Thrombolysis Decisions for Black and White Patients.” *Journal of General Internal Medicine* 22 (9): 1231–38.
Greenwald, A. G., D. E. McGhee, and J. L. K. Schwartz. 1998. “Measuring Individual Differences in Implicit Cognition: The Implicit Association Test.” *Journal of Personality and Social Psychology* 74 (6): 1464–80.
Gregg, A. P., B. Seibt, and M. R. Banaji. 2006. “Easier Done than Undone: Asymmetry in the Malleability of Implicit Preferences.” *Journal of Personality and Social Psychology* 90 (1): 1–20.

Hahn, A., C. M. Judd, H. K. Hirsh, and I. V. Blair. 2014. “Awareness of Implicit Attitudes.” *Journal of Experimental Psychology: General* 143 (3): 1369–92.
Hasson, U., and S. Glucksberg. 2006. “Does Understanding Negation Entail Affirmation?” *Journal of Pragmatics* 38 (7): 1015–32.
Heider, F. 1958. *The Psychology of Interpersonal Relations*. Hillsdale, New Jersey, London: Lawrence Erlbaum Associates Publishers.
Hieronymi, P. 2008. “Responsibility for Believing.” *Synthese* 161 (3): 357–73.
———. 2009. “Two Kinds of Agency.” In *Mental Actions*, edited by L. O’Brien and M. Soteriou. Oxford; New York: Oxford University Press.
Holroyd, J. 2012. “Responsibility for Implicit Bias.” *Journal of Social Philosophy* 43 (3): 274–306.
———. 2015. “Implicit Bias, Awareness and Imperfect Cognitions.” *Consciousness and Cognition* 33 (May): 511–23.
Holroyd, J. and D. Kelly. 2016. “Implicit Bias, Character, and Control.” In *From Personality to Virtue,* edited by A. Masala and J. Webber, Oxford: Oxford University Press.
Holroyd, J. and J. Sweetman. 2016. “The Heterogeneity of Implicit Bias.” In *Implicit Bias and Philosophy*, edited by M. Brownstein and J. Saul, Oxford: Oxford University Press.
Hume, D. 1748/1977. *An Enquiry Concerning Human Understanding*. Indianapolis: Hackett Publishing.
Hyman, J. 2015. *Action, Knowledge, and Will*. New York, NY: Oxford University Press.
Joel, D., Z. Berman, I. Tavor, N. Wexler, O. Gaber, Y. Stein, N. Shefi, et al. 2015. “Sex beyond the Genitalia: The Human Brain Mosaic.” *Proceedings of the National Academy of Sciences* 112 (50): 15468–73.
Kang, J., M. Bennett, D. Carbado, P. Casey, and D. Nilanjana. 2012. “Implicit Bias in the Courtroom.” *UCLA Law Review* 59: 1124–86.
Kant, I. 1781/2009. *Critique of Pure Reason*. Edited by P. Guyer, Translated by A. W. Wood. The Cambridge Edition of the Works of Immanuel Kant. Cambridge: Cambridge University Press.
Kelly, D., and E. Roedder. 2008. “Racial Cognition and the Ethics of Implicit Bias.” *Philosophy Compass* 3 (3): 522–40.
Kenny, A. 1963. *Action, Emotion, and Will*. London ; New York: Routledge.
Lane, K. A., M. R. Banaji, B. A. Nosek, and A. G. Greenwald. 2007. “Understanding and Using the Implicit Association Test: IV; What We Know (So Far) about the Method.” In *Implicit Measures of Attitudes*, edited by B. Wittenbrink and N. Schwarz. New York: Guilford Press.
Levy, N. 2005. “Libet’s Impossible Demand.” *Journal of Consciousness Studies* 12 (12): 67–76.
———. 2013. “The Importance of Awareness.” *Australasian Journal of Philosophy* 91 (2): 211–29.
———. 2014a. *Consciousness and Moral Responsibility*. Oxford ; New York: Oxford University Press.
———. 2014b. “Consciousness, Implicit Attitudes and Moral Responsibility.” *Noûs* 48 (1): 21–40.
———. 2015. “Neither Fish nor Fowl: Implicit Attitudes as Patchy Endorsements.” *Noûs* 49 (4): 800–823.
Lowery, B. S., C. D. Hardin, and S. Sinclair. 2001. “Social Influence Effects on Automatic Racial Prejudice.” *Journal of Personality and Social Psychology* 81 (5): 842–55.

Mandelbaum, E. forthcoming. “Attitude, Inference, Association: On the Propositional Structure of Implicit Bias: Attitude, Inference, Association.” *Noûs*.
McConnell, A. R., and J. M. Leibold. 2001. “Relations among the Implicit Association Test, Discriminatory Behavior, and Explicit Measures of Racial Attitudes.” *Journal of Experimental Social Psychology* 37 (5): 435–42.
Mele, A. R. 1992. *Springs of Action: Understanding Intentional Behavior*. New York: Oxford University Press.
Mele, A. R., and Paul K. Moser. 1994. “Intentional Action.” *Noûs* 28 (1): 39.
Mendoza, S. A., P. M. Gollwitzer, and D. M. Amodio. 2010. “Reducing the Expression of Implicit Stereotypes: Reflexive Control Through Implementation Intentions.” *Personality and Social Psychology Bulletin* 36 (4): 512–23.
Meyer, D. E., and R. W. Schvaneveldt. 1971. “Facilitation in Recognizing Pairs of Words: Evidence of a Dependence between Retrieval Operations.” *Journal of Experimental Psychology* 90 (2): 227–34.
Mill, J. S. 1843/2002. *A System of Logic: Ratiocinative and Inductive; Being a Connected View of the Principles of Evidence and the Methods of Scientific Investigation*. Hawaii: University of the Pacific.
Monteith, M. J., C. I. Voils, and L. Ashburn-Nardo. 2001. “Taking a Look Underground: Detecting, Interpreting, and Reacting to Implicit Racial Biases.” *Social Cognition* 19 (4): 395–417.
Moors, A., A. Spruyt, and J. De Houwer. 2010. “In Search of a Measure That Qualifies as Implicit: Recommendations Based on a Decompositional View of Automaticity.” In *Handbook of Implicit Social Cognition: Measurement, Theory, and Applications*, edited by B. Gawronski and B. K. Payne. New York: Guilford Press.
Moskowitz, G. B., and P. Li. 2011. “Egalitarian Goals Trigger Stereotype Inhibition: A Proactive Form of Stereotype Control.” *Journal of Experimental Social Psychology* 47 (1): 103–16.
Neely, J. H. 1977. “Semantic Priming and Retrieval from Lexical Memory: Roles of Inhibitionless Spreading Activation and Limited-Capacity Attention.” *Journal of Experimental Psychology: General* 106 (3): 226–54.
Nier, J. A. 2005. “How Dissociated Are Implicit and Explicit Racial Attitudes? A Bogus Pipeline Approach.” *Group Processes & Intergroup Relations* 8 (1): 39–52.
Nosek, B. A., A. G. Greenwald, and M. R. Banaji. 2005. “Understanding and Using the Implicit Association Test: II. Method Variables and Construct Validity.” *Personality and Social Psychology Bulletin* 31 (2): 166–80. doi:10.1177/0146167204271418.
Nosek, B. A., and M. R. Banaji. 2001. “The Go/No-Go Association Task.” *Social Cognition* 19 (6): 625–66.
Nosek, B. A., A. G. Greenwald, and M. R. Banaji. 2007. “The Implicit Association Test at Age 7: A Methodological and Conceptual Review.” In *Automatic Processes in Social Thinking and Behaviour*, edited by J. A. Bargh, 265–92. Psychology Press.
O’Connor, T. 2005. “Free Will.” In *Free Will: Critical Concepts in Philosophy*, edited by John Martin Fischer. Critical Concepts in Philosophy. London ; New York: Routledge.
Olson, M. A., and R. H. Fazio. 2008. “Implicit and Explicit Measures of Attitudes: The Perspective of the MODE Model.” In *Attitudes: Insights from*

the New Implicit Measures, edited by R. E. Petty, R. H. Fazio, and P. Briñol, 19–63. New York, NY, US: Psychology Press.
Payne, B. K. 2005. "Conceptualizing Control in Social Cognition: How Executive Functioning Modulates the Expression of Automatic Stereotyping." *Journal of Personality and Social Psychology* 89 (4): 488–503.
———. 2006. "Weapon Bias: Split-Second Decisions and Unintended Stereotyping." *Current Directions in Psychological Science* 15 (6): 287–91.
Payne, B. K., and B. Gawronski. 2010. "A History of Implicit Social Cognition: Where Is It Coming From? Where Is It Now? Where Is It Going?" In *Handbook of Implicit Social Cognition: Measurement, Theory, and Applications*, edited by B. Gawronski and B. K. Payne. New York: Guilford Press.
P. Briñol, R. E. Petty, and M. J. McCaslin. 2008. "Changing Attitudes on Implicit versus Explicit Measures: What Is the Difference?" In *Attitudes: Insights from the New Implicit Measures*, edited by R. E. Petty, R. H. Fazio, and P. Briñol, 19–63. New York, NY, US: Psychology Press.
Peters, K. R., and B. Gawronski. 2011. "Are We Puppets on a String? Comparing the Impact of Contingency and Validity on Implicit and Explicit Evaluations." *Personality and Social Psychology Bulletin* 37 (4): 557–69.
Petty, R. E., and P. Briñol. 2006. "A Metacognitive Approach to 'Implicit' and 'Explicit' Evaluations: Comment on Gawronski and Bodenhausen (2006)." *Psychological Bulletin* 132 (5): 740–44.
Petty, R. E., P. Briñol, and K. G. DeMarree. 2007. "The Meta–Cognitive Model (MCM) of Attitudes: Implications for Attitude Measurement, Change, and Strength." *Social Cognition* 25 (5): 657–86.
Plant, E. A., and P. G. Devine. 1998. "Internal and External Motivation to Respond without Prejudice." *Journal of Personality and Social Psychology* 75 (3): 811–32.
Plant, E. A., and B. M. Peruche. 2005. "The Consequences of Race for Police Officers' Responses to Criminal Suspects." *Psychological Science* 16 (3): 180–83.
Pollard, B. 2003. "Can Virtuous Actions Be Both Habitual and Rational?" *Ethical Theory and Moral Practice* 6 (4): 411–25.
Posner, M.I., and C. R. Snyder. 1975. "Attention and Cognitive Control." In *Information Processing and Cognition: The Loyola Symposium*, edited by R. L. Solso. Hillsdale, N.J.: New York: L. Erlbaum Associates.
Rachlinski, J. J., S. L. Johnson, A. J. Wistrich, and C. Guthrie. 2009. "Does Unconscious Bias Affect Trial Judges?" *Notre Dame Law Review* 84 (3): 1195–1246.
Rooth, D-O. 2007. "Implicit Discrimination in Hiring: Real World Evidence." *(IZA Discussion Paper No. 2764). Bonn, Germany: Forschungsinstitut Zur Zukunft Der Arbeit (Institute for the Study of Labor)*.
Rothermund, K., S. Teige-Mocigemba, A. Gast, and D. Wentura. 2009. "Minimizing the Influence of Recoding in the Implicit Association Test: The Recoding-Free Implicit Association Test (IAT-RF)." *Quarterly Journal of Experimental Psychology (2006)* 62 (1): 84–98.
Rozin, P., M. Markwith, and B. Ross. 1990. "The sympathetic magical law of similarity, nominal realism and neglect of negatives in response to negative labels." *Psychological Science* 1 (6): 383–84.
Rozin, P., L. Millman, and C. Nemeroff. 1986. "Operation of the Laws of Sympathetic Magic in Disgust and Other Domains." *Journal of Personality and Social Psychology* 50 (4): 703–12.

Rudman, L. A., and R. D. Ashmore. 2007. “Discrimination and the Implicit Association Test.” *Group Processes Intergroup Relations* 10 (3): 359–72.
Russell, B. 1912. *The Problems of Philosophy*. London: Hazen Press.
Sandis, C. 2015. “Verbal Reports and ‘Real’ Reasons: Confabulation and Conflation.” *Ethical Theory and Moral Practice* 18 (2): 267–80.
Saul, J. 2013. “Implicit Bias, Stereotype Threat and Women in Philosophy.” In *Women in Philosophy: What Needs to Change?*, edited by K. Hutchison and F. Jenkins. New York, NY: Oxford University Press.
Scanlon, T. M. 1998. *What We Owe to Each Other*. Cambridge, Mass.: Belknap Press of Harvard University Press.
Schmoldt, A., H. F. Benthe, and G. Haberland. 1975. “Digitoxin Metabolism by Rat Liver Microsomes.” *Biochemical Pharmacology* 24 (17): 1639–41.
Schneider, W., and R. M. Shiffrin. 1977. “Controlled and Automatic Human Information Processing: I. Detection, Search, and Attention.” *Psychological Review* 84 (1): 1–66.
Sher, G. 2006. “Out of Control.” *Ethics* 116 (2): 285–301.
Shoemaker, S. S. 1968. “Self-Reference and Self-Awareness.” *The Journal of Philosophy* 65 (19): 555–68.
Sloman, S. A. 1996. “The Empirical Case for Two Systems of Reasoning.” *Psychological Bulletin* 119 (1): 3–22.
Smith, A. M. 2005. “Responsibility for Attitudes: Activity and Passivity in Mental Life.” *Ethics* 115 (2): 236–71.
———. 2008. “Control, Responsibility, and Moral Assessment.” *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition* 138 (3): 367–92.
———. 2012. “Attributability, Answerability, and Accountability: In Defense of a Unified Account.” *Ethics* 122 (3): 575–89.
———. manuscript. “Implicit Biases, Moral Agency, and Moral Responsibility.”
Snow, N. E. 2006. “Habitual Virtuous Actions and Automaticity.” *Ethical Theory and Moral Practice* 9 (5): 545–61.
Sripada, C. S. 2010. “The Deep Self Model and Asymmetries in Folk Judgments about Intentional Action.” *Philosophical Studies* 151 (2): 159–76.
Steffens, M. C. 2004. “Is the Implicit Association Test Immune to Faking?” *Experimental Psychology* 51 (3): 165–79.
Stewart, B. D., and B. K. Payne. 2008. “Bringing Automatic Stereotyping Under Control: Implementation Intentions as Efficient Means of Thought Control.” *Personality and Social Psychology Bulletin* 34 (10): 1332–45.
Strawson, G. 2003. “XI-Mental Ballistics or The Involuntariness of Spontaneity.” *Proceedings of the Aristotelian Society (Hardback)* 103 (1): 227–56.
Strawson, P. F. 1962. “Freedom and Resentment.” Proceedings of the British Academy 48: 1–25.
Teige-Mocigemba, S., K. Christoph Klauer, and K. Rothermund. 2008. “Minimizing Method-Specific Variance in the IAT: A Single Block IAT.” *European Journal of Psychological Assessment* 24 (4): 237–45.
Uhlmann, E., and G. L. Cohen. 2005. “Constructed Criteria: Redefining Merit to Justify Discrimination.” *Psychological Science* 16 (6): 474–80.
van de Poel, I. 2011. “The Relation Between Forward-Looking and Backward-Looking Responsibility.” In *Moral Responsibility: Beyond Free Will and Determinism*, edited by N. Vincent, I. van de Poel, and J. Hoven, 37–52. Dordrecht: Springer Netherlands.
Velleman, D. 1992. “What Happens When Someone Acts?” *Mind* 101 (403): 461–81.

———. 2000. “On the Aim of Belief.” In *The Possibility of Practical Reason*, edited by D. Velleman, 244–81. Oxford University Press.
Washington, N., and D. Kelly. 2016. “Who’s Responsible for This? Moral Responsibility, Externalism, and Knowledge about Implicit Bias.” In *Implicit Bias and Philosophy*, edited by M. Brownstein and J. Saul. New York: Oxford University Press.
Watson, G. 1975. “Free Agency.” *The Journal of Philosophy* 72 (8): 205–20.
———. 1996. “Two Faces of Responsibility.” *Philosophical Topics* 24 (2): 227–48.
Webb, T. L., and P. Sheeran. 2006. “Does Changing Behavioral Intentions Engender Behavior Change? A Meta-Analysis of the Experimental Evidence.” *Psychological Bulletin* 132 (2): 249–68.
———. 2008. “Mechanisms of Implementation Intention Effects: The Role of Goal Intentions, Self-Efficacy, and Accessibility of Plan Components.” *British Journal of Social Psychology* 47 (3): 373–95.
Webb, T. L., P. Sheeran, and J. Pepper. 2012. “Gaining Control over Responses to Implicit Attitude Tests: Implementation Intentions Engender Fast Responses on Attitude-Incongruent Trials.” *The British Journal of Social Psychology / the British Psychological Society* 51 (1): 13–32.
Williams, B. 1973. *Problems of the Self; Philosophical Papers 1956-1972*. Cambridge: Cambridge University Press.
Wilson, T. D., S. Lindsey, and T. Y. Schooler. 2000. “A Model of Dual Attitudes.” *Psychological Review* 107 (1): 101–26.
Wolf, S. 1990. *Freedom within Reason*. New York: Oxford Univ. Press.